



THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

Appellant: Madison *et al.*
Appl. No.: 09/776,191
Conf. No.: 3237
Filed: February 2, 2001
Title: **NUCLEIC ACID MOLECULES ENCODING TRANSMEMBRANE
SERINE PROTEASES, THE ENCODED PROTEINS AND METHODS
BASED THEREON**
Art Unit: 1652
Examiner: Yong D. Pak

Mail Stop Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPELLANT'S APPEAL BRIEF

Sir:

Appellant submits this Appeal Brief in support of the Notice of Appeal, filed on August 14, 2008. This Appeal is from the Final Rejection in the Office Action, dated March 26, 2008. The Appeal Brief is filed with a five-month Extension of Time under Rule 136(a).

03/18/2009 WABDELRI 00000073 021818 09776191

01 FC:2402 270.00 DA

CERTIFICATE OF MAILING BY "EXPRESS MAIL"
"Express Mail" Mailing Label Number EV 740126652 US
Date of Deposit: **March 16, 2009**

I hereby certify that this paper is being deposited with the United States Postal "Express Mail Post Office to Addressee" Service under 37 CFR §1.10 on the date indicated above and is addressed to: Mail Stop Appeal Brief-Patents, Commissioner for Patents, U.S. Patent and Trademark Office, P.O. Box 1450, Alexandria, VA, 22313-1450.


Frank J. Miskiel

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

I. REAL PARTY IN INTEREST

The real party in interest for the above-identified patent application on Appeal is
Dendreon Corporation
by virtue of an Assignment recorded May 20, 2002 at reel 014703, frame 0441 in the United
States Patent and Trademark Office.

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

II. RELATED APPEALS AND INTERFERENCES

Appellant's legal representative and the Assignee of the above-identified patent application do not know of any prior or pending appeals, interferences or judicial proceedings that may be related to, directly affect or be directly affected by or have a bearing on the Board's decision with respect to the above-identified Appeal.

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

III. STATUS OF CLAIMS

Claims 1, 10-13, 20, 34-36, 40-46, 48-55, 108, 109, 113-116, 118-120 and 122-126 are pending in the above-identified patent application. Claims 10, 43-46, 48-55, 108, 109, 115, 116, 118-120 and 122-126 are withdrawn from consideration, but are retained for possible rejoinder upon allowance of a generic claim. Claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 are rejected. Therefore, Claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 are the subject of this appeal. A copy of the appealed claims, and all pending claims, is included in the Claims Appendix.

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

IV. STATUS OF AMENDMENTS

No amendment was filed subsequent to the final rejection. Appellant filed a Notice of Appeal on August 14, 2008 (mailed on that date via Express mail certificate of mailing).

Appellant attaches a copy of the Final Office Action as Exhibit 1 in the Evidence Appendix.

9

V. SUMMARY OF CLAIMED SUBJECT MATTER

The following is a brief discussion of subject matter of the claimed subject matter. As described and defined in the application (see, *e.g.*, page 7, last paragraph- page 8; and page 18, line 13, - page 19). Transmembrane serine protease (hereinafter MTSPs) are a known family of serine proteases. Their identity and sequences are known, and, the prior art teaches that these proteases require activation and cleavage for activity. The active form is typically a two chain or other multi-chain form. There is no teaching or suggestion in any art, that isolated protease domains of the protease as a single chain has activity, nor is there any teaching or suggestion for isolating such domain. Independent claim 1 is directed to isolated single chain protease domains of an MTSP that are modified by replacing a free cysteine with another amino acid; all claims are dependent thereon. The free cysteine in the protease domain, is not free in the activated full-length molecule. Modification of the single chain protease domain by replacing the free cysteine prevents aggregation that occurs by virtue of interaction among the free cysteines among molecules. Since none of the art suggests that the isolated protease domain has activity, none can suggest modifying the isolated protease domain to avoid aggregation which will impact on activity.

As defined in the application (pages 18-20), an MTSP family member is:

As used herein, "transmembrane serine protease (MTSP)" refers to a family of transmembrane serine proteases that share common structural features as described herein (see, also Hooper *et al.* (2001) *J. Biol. Chem.* 276:857-860). Thus, reference, for example, to "MTSP" encompasses all proteins encoded by the MTSP gene family, including but are not limited to: MTSP1, MTSP3, MTSP4 and MTSP6, or an equivalent molecule obtained from any other source or that has been prepared synthetically or that exhibits the same activity. Other MTSPs include, but are not limited to, corin, enteropeptidase, human airway trypsin-like protease (HAT), MTSP1, TMPRSS2, and TMPRSS4. Sequences of encoding nucleic molecules and the encoded amino acid sequences of exemplary MTSPs and/or domains thereof are set forth in SEQ ID Nos. 1-12, 49, 50 and 61-72. The term also encompass MTSPs with conservative amino acid substitutions that do not substantially alter activity of each member, and also encompasses splice variants thereof. Suitable conservative substitutions of amino acids are known to those of skill in this art and may be made generally without altering the biological activity of the resulting molecule. Of particular interest are MTSPs of mammalian, including human, origin. Those of skill in this art recognize that, in general, single amino acid substitutions in non-essential regions of a polypeptide do not substantially alter biological activity (see, *e.g.*, Watson *et al.* *Molecular Biology of the Gene*, 4th Edition, 1987, The Benjamin/Cummings Pub. Co., p. 224).

The application identifies the known members of the family: corin, enteropeptidase, human airway trypsin-like protease (HAT), hepsin, MTSP1, TMPRSS2, TMPRSS4 and TADG-12), and provides sequences of numerous family members and also provides new family members

(*e.g.*, MTSP3, MTSP4 and MTSP6). Pages 10-12 reference sequence identifiers and or references providing the sequences of each member of the family:

... corin (accession nos. AF133845 and AB013874; see, Yan *et al.* (1999) *J. Biol. Chem.* 274:14926-14938; Tomita *et al.* (1998) *J. Biochem.* 124:784-789; Uan *et al.* (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97:8525-8529; SEQ ID Nos. 61 and 62 for the human protein); enteropeptidase (also designated enterokinase; accession no. U09860 for the human protein; see, Kitamoto *et al.* (1995) *Biochem.* 27: 4562-4568; Yahagi *et al.* (1996) *Biochem. Biophys. Res. Commun.* 219:806-812; Kitamoto *et al.* (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91:7588-7592; Matsushima *et al.* (1994) *J. Biol. Chem.* 269:19976-19982; see SEQ ID Nos. 63 and 64 for the human protein); human airway trypsin-like protease (HAT; accession no. AB002134; see Yamaoka *et al.* *J. Biol. Chem.* 273:11894-11901; SEQ ID Nos. 65 and 66 for the human protein); hepsin (see, accession nos. M18930, AF030065, X70900; Leytus *et al.* (1988) *Biochem.* 27: 11895-11901; Vu *et al.* (1997) *J. Biol. Chem.* 272:31315-31320; and Farley *et al.* (1993) *Biochem. Biophys. Acta* 1173:350-352; SEQ ID Nos. 67 and 68 for the human protein); TMPRS2 (see, Accession Nos. U75329 and AF113596; Paoloni-Giacobino *et al.* (1997) *Genomics* 44:309-320; and Jacquinet *et al.* (2000) *FEBS Lett.* 468: 93-100; SEQ ID Nos. 69 and 70 for the human protein) TMPRSS4 (see, Accession No. NM 016425; Wallrapp *et al.* (2000) *Cancer* 60:2602-2606; SEQ ID Nos. 71 and 72 for the human protein); and TADG-12 (also designated MTSP6, see SEQ ID Nos. 11 and 12; see International PCT application No. WO 00/52044, which claims priority to U.S. application Serial No. 09/261,416).

... Exemplary MTSPs (see, *e.g.*, SEQ ID No. 1-12, 49 and 50) are provided herein, as are the single chain protease domains thereof as follows: SEQ ID Nos. 1, 2, 49 and 50 set forth amino acid and nucleic acid sequences of MTSP1 and the protease domain thereof; SEQ ID No. 3 sets forth the MTSP3 nucleic acid sequence and SEQ ID No. 4 the encoded MTSP3 amino acids; SEQ ID No. 5 MTSP4 a nucleic acid sequence of the protease domain and SEQ ID No. 6 the encoded MTSP4 amino acid protease domain; SEQ ID No. 7 MTSP4-L a nucleic acid sequence and SEQ ID No. 8 the encoded MTSP4-L amino acid sequence; SEQ ID No. 9 an MTSP4-S encoding nucleic acid sequence and SEQ ID No. 10 the encoded MTSP4-S amino acid sequence; and SEQ ID No. 11 an MTSP6 encoding nucleic acid sequence and SEQ ID No. 12 the encoded MTSP6 amino acid sequence. The single chain protease domains of each are delineated below.

As described in the application, and noted above, Appellant has discovered that the protease domain as a single chain polypeptide that contains only the protease domain of an MTSP protease possesses protease activity. Prior to this the dogma in the protease field was that these serine proteases exist as a zymogen that requires activation cleavage for activity. Activation cleavage cleaves the disulfide bond that forms between a cysteine residue in the protease domain and another domain of the enzyme. As a result of the activation cleavage, the active protease occurs as a two-chain or multi-chain molecule. See, *e.g.*, Lin *et al.*, (*J. Biol. Chem.* 274:18231-18236 (1999), Exhibit 20, which teaches that serine proteases are synthesized as single-chain zymogens, which are proteolytically activated to become active two-chain forms (*e.g.*, see page 18235, col. 2, first full paragraph); and Takeuchi *et al.* (*Proc.*

Natl. Acad. Sci. USA 96: 11054-11061 (1999), Exhibit 3), which describes the pro-domain region of its MTSP1 as disulfide bonded to the protease domain (see page 11058, col. 1 and page 11060, col. 1, first paragraph) and remains bonded to the protease domain after auto-activation (page 11058, lines 8-9), resulting in a polypeptide that includes a protease domain disulfide bonded to a pro-domain having a two-chain form.

The application teaches (see, *e.g.*, page 8, lines 15-21; page 20, lines 1-6; page 25, line 4 through page 26, line 25; page 58, lines 5-11) that the single chain protease domain is active. The application also teaches how to identify a protease domain (see, *e.g.*, page 8, lines 7-14 and page 19, lines 3-24). For example, at page 18, line 24 through page 20, line 6, the specification defines a protease domain of an MTSP as well as the requisites for activity and how to identify a protease domain as:

As used herein, a "protease domain of an MTSP" refers to the protease domain of MTSP that is located within the extracellular domain of a MTSP and exhibits serine proteolytic activity. It includes at least the smallest fragment thereof that acts catalytically as a single chain form. Hence it is at least the minimal portion of the extracellular domain that exhibits proteolytic activity as assessed by standard assays *in vitro* assays. Those of skill in this art recognize that such protease domain is the portion of the protease that is structurally equivalent to the trypsin or chymotrypsin fold.

Exemplary MTSP proteins, with the protease domains indicated, are illustrated in Figures 1-3. Smaller portions thereof that retain protease activity are contemplated. The protease domains vary in size and constitution, including insertions and deletions in surface loops. **They retain conserved structure, including at least one of the active site triad, primary specificity pocket, oxyanion hole and/or other features of serine protease domains of proteases.** Thus, for purposes herein, the protease domain is a portion of a MTSP, as defined herein, and is homologous to a domain of other MTSPs, such as corin, enteropeptidase, human airway trypsin-like protease (HAT), MTSP1, TMPRSS2, and TMPRSS4, which have been previously identified; it was not recognized, however, that an isolated single chain form of the protease domain could function proteolytically in *in vitro* assays. **As with the larger class of enzymes of the chymotrypsin (S1) fold (see, *e.g.*, Internet accessible MEROPS data base), the MTSPs protease domains share a high degree of amino acid sequence identity. The His, Asp and Ser residues necessary for activity are present in conserved motifs.** The activation site, which results in the N-terminus of second chain in the two chain forms is has a conserved motif and readily can be identified (see, *e.g.*, amino acids 801-806, SEQ ID No. 62, amino acids 406-410, SEQ ID No. 64; amino acids 186-190, SEQ ID No. 66; amino acids 161-166, SEQ ID No. 68; amino acids 255-259, SEQ ID No. 70; amino acids 190-194, SEQ ID No. 72).

As used herein, the catalytically active domain of an MTSP refers to the protease domain. . . .

Significantly, it is shown herein, that, at least *in vitro*, the single chain forms of the MTSPs and the catalytic domains or proteolytically active portions thereof (typically C-terminal truncations) thereof exhibit protease activity. Hence provided herein are isolated single chain forms of the protease

domains of MTSPs and their use in *in vitro* drug screening assays for identification of agents that modulate the activity thereof.

The specification teaches modified protease domains (see, *e.g.*, page 11, the description for each of the working examples, and the working examples, which describe replacement of the free (unpaired) Cys residue in the protease domain):

Also provided are muteins of the single chain protease domains and MTSPs, particularly muteins in which the Cys residue in the protease domain that is free (*i.e.*, does not form disulfide linkages with any other Cys residue in the protein) is substituted with another amino acid substitution, preferably with a conservative amino acid substitution or a substitution that does not eliminate the activity, and muteins in which a glycosylation site(s) is eliminated. Muteins in which other conservative amino acid substitutions in which catalytic activity is retained are also contemplated (see, *e.g.*, Table 1, for exemplary amino acid substitutions). See, also, Figure 4, which identifies the free Cys residues in MTSP3, MTSP4 and MTSP6.

CLAIMS ON APPEAL AND EXEMPLARY SUPPORTING DISCLOSURE IN THE APPLICATION

Claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 are the subject of this appeal and each is argued separately throughout. Independent Claim 1 is directed to an isolated, substantially purified (*e.g.*, see page 46, lines 4-15) single-chain polypeptide, **consisting only** of a protease domain of a type-II membrane-type serine protease (MTSP) (*e.g.*, see page 17, line 24 through page 19, line 2 and page 25, line 4-page 26, line 12) or a catalytically active fragment thereof (*e.g.*, see page 26, lines 13-25) as a single chain (*e.g.*, see page 26, lines 13-25 and 58, lines 5-11), wherein a **free Cys** (*e.g.*, see page 10, lines 4-6) in the protease domain is replaced with another amino acid (*e.g.*, see page 10, lines 3-13); and the MTSP protease domain or catalytically active fragment thereof has serine protease activity (*e.g.*, see page 31, lines 14-20) as a single chain (*e.g.*, see page 26, lines 13-25 and 58, lines 5-20; original claim 1). All claims ultimately depend from claim 1.

Dependent claim 11 is directed to the substantially purified polypeptide of claim 1, wherein the MTSP is selected from among MTSP1, MTSP3, MTSP4 and MTSP6 (*e.g.*, see page 8, line 30 through page 9, line 8 and original claim 11).

Dependent claim 12 is directed to the substantially purified (*e.g.*, see page 46, lines 4-15) polypeptide of claim 1, where the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12 (*e.g.*, see page 25, lines 22-27 and original claim 12).

Dependent claim 13 is directed to the substantially purified (*e.g.*, see page 46, lines 4-15) polypeptide of claim 1 that has at least about 95% sequence identity with a protease

domain consisting of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acids set forth as SEQ ID No. 6, and amino acids 217-443 in SEQ ID No. 12 (*e.g.*, see page 25, lines 22-31 and original claim 13).

Dependent claim 20 is directed to the polypeptide of claim 1, where a free Cys in the protease domain is replaced with a serine (*e.g.*, see page 10, lines 3-13, page 163, lines 4-8 and original claim 20).

Dependent claim 34 is directed to the polypeptide of claim 1, where the MTSP is selected from among corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4 (*e.g.*, see page 8, line 30 through page 9, line 8 and original claim 34).

Dependent claim 35 is directed to a conjugate (*e.g.*, see page 38, lines 1-8 and page 123, line 30 through page 136, line 2), that includes a) a polypeptide of claim 1, and b) a targeting agent (*e.g.*, see page 38, lines 9-15 and page 130, lines 9-17) linked to the protein directly or via a linker (*e.g.*, see page 126, line 9 through page 130, line 7), where the conjugate has serine protease activity (*e.g.*, see page 10, lines 3-13 and original claim 35).

Dependent claim 36 is directed to a conjugate of claim 35, wherein the targeting agent permits i) affinity isolation or purification of the conjugate; ii) attachment of the conjugate to a surface; iii) detection of the conjugate; or iv) targeted delivery to a selected tissue or cell (*e.g.*, see page 14, lines 19-26 and original claim 36).

Dependent claim 40 is directed to a solid support (*e.g.*, see page 126, lines 12-15) comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker (*e.g.*, see page 131, line 92 through page 134, line 30 and original claims 39).

Dependent claim 41 is directed to the solid support of claim 40 and recites that the polypeptides comprise an array (*e.g.*, see page 132, lines 4-8 and original claim 40).

Dependent claim 42 is directed to the solid support of claim 41 and recites that the array includes polypeptides having different MTSP protease domains (*e.g.*, see and original claim 41).

Dependent claim 113 is directed to a solid support (*e.g.*, see page 126, lines 12-15) comprising two or more polypeptides of claim 12 linked thereto either directly or via a linker (*e.g.*, see page 126, line 9 through page 130, line 7 and original claim 112).

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

Claim 114 depends from claim 113 and specifies that the polypeptides comprise an array (*e.g.*, see page 132, lines 4-8 and original claim 113).

A list of the currently pending claims is provided in the Claims Appendix of this Brief.

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. Rejections under 35 U.S. C. § 112, first paragraph

1. Claims 1, 11, 20, 34-36, 40-42, 113 and 114 are rejected under 35 U.S.C. §112, first paragraph, as containing subject matter that was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed subject matter.
2. Claims 1, 11, 20, 34-36, 40-42, 113 and 114 are rejected under 35 U.S.C. §112, first paragraph, because the specification, while being enabling for a polypeptide consisting of amino acids 615-855 of SEQ ID NO:2, allegedly does not reasonably provide enablement for a polypeptide consisting of any protease domain of any type II membrane type serine protease (MTSP) or a catalytically active portion thereof.

B. Rejection under 35 U.S.C. 102(b)

Claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 are rejected under 35 U.S.C. §102(b) as being anticipated by Takeuchi *et al.*, Proc. Natl. Acad. Sci. USA 96: 11054-11061 (1999) ("Takeuchi"), a copy of which is attached in the Evidence Appendix as Exhibit 3.

C. Rejection under 35 U.S.C. 102(e)

Claims 1, 11-13 and 34 are rejected under 35 U.S.C. §102(e) as anticipated by O'Brien *et al.*, U.S. Patent No. 5,972,616 ("O'Brien"), a copy of which is attached in the Evidence Appendix as Exhibit 4.

D. Rejection under 35 U.S.C. 103(a)

Claims 1, 11-13 and 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 103(a) as being unpatentable over O'Brien.

VII. ARGUMENTS

1. REJECTION OF CLAIMS 1, 11, 20, 34-36, 40-42, 113 AND 114 UNDER 35 U.S.C. §112, FIRST PARAGRAPH – POSSESSION

Claims 1, 11, 20, 34-36, 40-42, 113 and 114 are rejected under 35 U.S.C. §112, first paragraph, as allegedly containing subject matter that was not described in the specification in such a way as to reasonably convey to one skilled in the art that the inventor, at the time the application was filed, had possession of the claimed subject matter. The Examiner alleges that claims 1, 11, 20, 34-36, 40-42 and 113-114 are drawn to a polypeptide consisting of a protease domain or catalytically active fragment thereof of type-II membrane-type serine protease (MTSP) from any source and concludes that these claims are drawn to a genus of polypeptides having any structure. The Examiner alleges that the specification only teaches four species, and that four species are not a sufficient number of representative species of the genus to describe the whole genus. The Examiner also alleges that there is no evidence on the record of the relationship between the structure of the exemplary catalytically active protease domains and the structure of the serine protease domain of any or all MTSP polypeptides or MTSP1 polypeptides. The Final Office Action concludes that the specification fails to sufficiently describe the claimed subject matter in such full, clear, concise, and exact terms that a skilled artisan would recognize that Appellant was in possession of the claimed subject matter. The rejection respectfully is traversed.

A. LEGAL STANDARDS - 35 U.S.C. §112, FIRST PARAGRAPH – POSSESSION

The purpose behind the written description requirement is to ensure that the patent Appellant had possession of the claimed subject matter at the time of filing of the application. The relevant law and a discussion of the Patent Office Guidelines are set forth in the previous responses of record in this application and below. Briefly, the Federal Circuit has discussed the application of the written description requirement of the first paragraph of 112 to claims in the field of biotechnology. See *University of California v. Eli Lilly and Co.*, 119 F.3d 1559, 43 U.S.P.Q.2d 1398, 1406 (Fed. Cir. 1997). The court explained that:

In claims involving chemical materials, generic formulae usually indicate with specificity what the generic claims encompass. One skilled in the art can distinguish such a formula from others and can identify many of the species that the claims encompass. Accordingly, such a formula is normally an adequate description of the claimed genus . . . a generic statement such as "vertebrate insulin or "mammalian insulin without more, is

not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function. It does not specifically define any of the genes that fall within its definition. It does not define any structural features commonly possessed by members of the genus that distinguish them from others. One skilled in the art therefore cannot, as one can do with a fully described genus, visualize or recognize the identity of the members of the genus. A definition by function, as we have previously indicated, does not suffice to define the genus because it is only an indication of what the gene does, rather than what it is.

The court also stated that “[a]written description of an invention involving a chemical genus, like a description of a chemical species, ‘requires a precise definition, such as by structure, formula, [or]chemical name,’ of the claimed subject matter sufficient to distinguish it from other materials.” *Id.* at 1567, 43 U.S.P.Q.2d at 1405. Finally, the court addressed the manner by which a genus of might be described. “A description of a genus of cDNA may be achieved by means of a recitation of a representative number of cDNAs, defined by nucleotide sequence, falling within the scope of the genus or of a recitation of structural features common to the members of the genus, which features constitute a substantial portion of the genus.” *Id.*

The Federal Circuit also has addressed the written description requirement in the context of biotechnology-related subject matter in *Enzo Biochem. Inc. v. Gen-Probe*, 296 F.3d 1316, 63 USPQ2d (BNA) 1609 (Fed. Cir. 2002). The Enzo court adopted the standard that:

the written description requirement can be met by ‘showing that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics . . . complete or partial structure, other physical chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics.’

The court in *Enzo* adopted its standard from the Written Description Examination Guidelines. The Guidelines apply to proteins as well as nucleic acid molecules.

It is well-settled that the written description requirement of 35 U. S. C. §112, first paragraph, can be satisfied without express or explicit disclosure of a later-claimed invention. See, *In re Herschler*, 591 F.2d 693, 700-01, 200 USPQ 711, 717 (CCPA 1979):

"The claimed subject matter need not be described in *haec verba* to satisfy the description requirement. It is not necessary that the application describe the claim limitations exactly, but only so clearly that one having ordinary skill in the pertinent art would recognize from the disclosure that appellants invented processes including those limitations." (citations omitted).

See also *Purdue Pharma L. P. v. Faulding, Inc.*, 230 F.3d 1320, 56 USPQ2d 1481 (Fed. Cir. 2000).

The written description requirement of 35 U.S.C § 112, first paragraph, can be satisfied by providing sufficient disclosure, either through illustrative examples or terminology. This clause does not require "a specific example of everything within the scope of a broad claim." In *re* Anderson, 176 USPQ 331, at 333 (CCPA 1973), emphasis in original. Further, because "it is manifestly impracticable for an applicant who discloses a generic invention to give an example of every species falling within it, or even to name every such species, it is sufficient if the disclosure teaches those skilled in the art what the invention is and how to practice it." In *re* Grimme, Keil and Schmitz, 124 USPQ 449, 502 (CCPA 1960).

B. THE REJECTION OF CLAIMS 1-3, 5, 9, 11, 19, 20, 34-36, 40-42, 113 AND 114 UNDER 35 U.S.C. §112, FIRST PARAGRAPH SHOULD BE REVERSED BECAUSE THE SPECIFICATION MEETS THE WRITTEN DESCRIPTION REQUIREMENT WITH RESPECT TO POSSESSION

Claim 1

In setting forth the rejection, the Examiner states that the claims are drawn to polypeptides having any structure and are thus drawn to a genus encompassing species having substantial variation. The Examiner states that only four species are described in the specification and that there is no evidence on the record of the relationship between the structure of the exemplary catalytically active protease domains and the structure of the serine protease domain of any or all MTSP polypeptides. Appellant respectfully submits that this is not correct.

1. Standard for satisfying the written description requirement for possession

In order to satisfy the written description requirement, one need not provide an example of every species encompassed by a claim. It is sufficient to provide identifying characteristics, including structural and physical characteristics, functional characteristics coupled with known or disclosed correlation with structural characteristics to demonstrate that the applicant was in possession of the claimed subject matter. MPEP § 2163; see *University of California v. Eli Lilly*, 119 F. 3d 1559, 1568, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997). Further, the standard is an objective one, based on what one of skill in the art would recognize in the disclosure. *In re Gosteli*, 872 F.2d at 1012. As is discussed in more detail below, it respectfully is submitted that the instant application sufficiently describes the claimed genus of isolated MTSP protease domains to demonstrate possession of the claimed subject matter at the time of the effective filing date of each claim as required by this standard.

2. Specification describes more than four species of MTSP protease domains

In this instance, the specification identifies all known members of the family and identifies several new members, including protease domains (as well as full-length) MTSP3, MTSP6 two splice variants of MTSP4. Thus, contrary to the Examiner's assertion that the specification provides only four species of protease domains, Appellant respectfully submits that the application **identifies all of the 17 known members of the MTSP family** (see, *e.g.*, page 4) known at the time of filing, and provides the sequences of full-length MTSP proteases and identifies the protease domains thereof. In addition, the specification teaches how to identify a protease domain in an MTSP, how to identify a free Cys residue and to replace a Cys residue. The members of the MTSP family provided include, MTSP1 (also referred to as matriptase and TAGD-15), MTSP3, MTSP4 (two variants encoded by splice variants), MTSP6, corin, enteropeptidase, human airway trypsin-like protease (HAT), hepsin, TMPRS2 and TMPRSS4. For example, page 4, line 20 through page 5, line 17 of the specification recites:

In mammals, at least 17 members of the family are known, including seven in humans (see, Hooper *et al.* (2001) J. Biol. Chem. 276:857-860). These include: corin (accession nos. AF133845 and AB013874; see, Yan *et al.* (1999) J. Biol. Chem. 274:14926-14938; Tomita *et al.* (1998) J. Biochem. 124:784-789; Uan *et al.* (2000) Proc. Natl. Acad. Sci. U.S.A. 97:8525-8529); enteropeptidase (also designated enterokinase; accession no. U09860 for the human protein; see, Kitamoto *et al.* (1995) Biochem. 27: 4562-4568; Yahagi *et al.* (1996) Biochem. Biophys. Res. Commun. 219:806-812; Kitamoto *et al.* (1994) Proc. Natl. Acad. Sci. U.S.A. 91:7588-7592; Matsushima *et al.* (1994) J. Biol. Chem. 269:19976-19982); human airway trypsin-like protease (HAT; accession no. AB002134; see Yamaoka *et al.* J. Biol. Chem. 273:11894-11901); MTSP1 and matriptase (also called TAGD-15; see SEQ ID Nos. 1 and 2; accession nos. AF133086/AF118224, AF04280022; Takeuchi *et al.* (1999) Proc. Natl. Acad. Sci. U.S.A. 96:11054-1161; Lin *et al.* (1999) J. Biol. Chem. 274:18231-18236; Takeuchi *et al.* (2000) J. Biol. Chem. 275:26333-26342; and Kim *et al.* (1999) Immunogenetics 49:420-429); hepsin (see, accession nos. M18930, AF030065, X70900; Leytus *et al.* (1988) Biochem. 27: 11895-11901; Vu *et al.* (1997) J. Biol. Chem. 272:31315-31320; and Farley *et al.* (1993) Biochem. Biophys. Acta 1173:350-352; and see, U.S. Pat. No. 5,972,616); TMPRS2 (see, Accession Nos. U75329 and AF113596; Paoloni-Giacobino *et al.* (1997) Genomics 44:309-320; and Jacquinet *et al.* (2000) FEBS Lett. 468: 93-100); and TMPRSS4 (see, Accession No. NM 016425; Wallrapp *et al.* (2000) Cancer 60:2602-2606).

Thus, the specification provides 17 examples of MTSPs and isolated protease domains (*e.g.*, see also pages 9-10), including MTSP1, MTSP3, MTSP4 (2 splice variants) and MTSP6, incorporates publications describing all known family members and the protease domains thereof, and describes full-length sequences.

3. MTSPs are a known family of serine proteases with known structural features

As noted, the MTSPs are a known and well studied family of enzymes, the specification teaches how to identify members of the MTSP family and the specification provides relevant structural and functional features that uniquely identify and specify the claimed genus of polypeptides. The MTSP protease family of enzymes has been extensively studied and characterized, evidenced by the art made of record in Information Disclosure Statements and provided in previous responses and herein. Hooper *et al.* teaches that many of the serine proteases are mosaic proteins that include multiple, structurally distinct domains necessary for regulating enzymatic activity (Eur. J. Biochem. 267: 6931-6937 (2000), Exhibit 14). Lin *et al.* ((1999) J. Biol. Chem. 274:18231-36, Exhibit 20) and Yan *et al.* ((1999) J. Biol. Chem. 274:14926-35), Exhibit 44) teach that MTSPs are a family of proteins that can be distinguished from many other types of proteins and enzymes because they have highly conserved structures. For example, as discussed in the instant specification, it is known in the art that a substrate specificity pocket in the protease domain and conserved cysteines that participate in disulfide bonding are highly conserved features in serine proteases (see, e.g., Figure 4 and page 18235 of Lin *et al.* (Exhibit 20) and Figure 2 and page 18236 of Yan *et al.*, Exhibit 44).

MTSPs are a class of serine proteases characterized by having an NH₂-terminal cytoplasmic tail and a COOH-terminal ectodomain, lacking an NH₂-terminal cleavable signal sequence, and having a signal/anchor domain that anchors the serine protease in the cell membrane (e.g., see Parks *et al.*, J. Biol. Chem. 268: 19101-19109 (1993), Exhibit 26 and Parks & Lamb, Cell 64: 777-787 (1991), Exhibit 27). Tsuji *et al.* teaches that MTSPs, such as hepsin, include a hydrophobic sequence flanked by a sequence having a positive net charges on the NH₂-terminal side while the COOH-terminal flanking side contains no charge, which agrees with the consensus topological sequence for the MTSPs (Tsuji *et al.*, J Biol Chem 266(25): 16948-16953 (1991), Exhibit 37). The MTSPs have the triad of residues His57, Asp102 and Ser195 at the active site (chymotrypsin numbering system), which are in close proximity and serve as a functional interacting unit responsible for bond formation and cleavage during catalysis (Craik *et al.*, Science 237:909-913 (1987), Exhibit 10). Thus, an MTSP polypeptide can be characterized as a serine protease that includes the conserved catalytic triad, lacks a cleavable signal sequence, includes a transmembrane anchoring domain, and has positively charged residues on the N-terminal side of a long stretch of hydrophobic amino acids and has a characteristic disulfide bond pattern (Walter *et al.*, Annu.

Rev. Cell Biol. 2: 499-516 (1986), Exhibit 40). The lack of a signal sequence, a characteristic disulfide bond pattern, a characteristic hydrophobic region and the presence of a signal/anchor domain also are seen in all of the MTSPs, including hepsin (Leytus *et al.*, Biochemistry 27: 1067-1074 (1988), Exhibit 19), enteropeptidase (Kitamoto *et al.*, Proc. Natl. Acad. Sci. USA 91: 7588-7592 (1994), Exhibit 17), TMPRSS2 (Paoloni-Giacobino *et al.*, Genomics 44: 309-320 (1997), Exhibit 31), and human airway trypsin-like protease (Yamaoka *et al.*, J. Biol. Chem. 273: 11895-11901 (1998), Exhibit 43).

The specification also describes structural features and structure-function relationships that identify the MTSP family of polypeptides. Such description includes information regarding the tertiary structure of the polypeptide. For example, the specification teaches the locus of the disulfide bonds, identifies the Cys residues that link the protease domain to the rest of the polypeptide, and teaches that the polypeptide includes at least one of the active site triad, primary specificity pocket and oxyanion hole. The specification states that the MTSP family of proteins shares a high degree of homology. Hence, other MTSPs, such as MTSPs from other species, can be readily identified by its homology with known MTSPs. The specification also teaches that the protease domain of a MTSP shares homology and structural features with the chymotrypsin/trypsin family protease domains. The previous responses of record and the application establish that the application describes the MTSP family and describes identification and isolation of protease domains.

Most significantly, the application identifies the known members of the MTSP family, provides sequences thereof and/or references earlier publications describing the family members, and provides working examples for MTSP1, MTSP3, MTSP6 and the two MTSP4 splice variants.

4. The specification provides relevant identifying characteristics of the protease domain

As discussed in responses of record, methods of identifying and isolating serine protease domains of MTSPs were known in the art at the time of filing the application and are taught in the specification. The specification describes protease domains of MTSPs and provides sequences of exemplars thereof. For example, the specification teaches, *e.g.*, at page 19, lines 3-24, that:

Exemplary MTSP proteins, with the protease domains indicated, are illustrated in Figures 1-3. Smaller portions thereof that retain protease activity are contemplated. The protease domains vary in size and constitution, including insertions and deletions in surface loops. They **retain conserved structure**, including at least one of the active

site triad, primary specificity pocket, oxyanion hole and/or other features of serine protease domains of proteases. Thus, for purposes herein, the protease domain is a portion of a MTSP, as defined herein, and is homologous to a domain of other MTSPs, such as corin, enteropeptidase, human airway trypsin-like protease (HAT), MTSP1, TMPRSS2, and TMPRSS4, which have been previously identified; it was not recognized, however, that an isolated single chain form of the protease domain could function proteolytically in *in vitro* assays. As with the larger class of enzymes of the chymotrypsin (S1) fold (see, e.g., Internet accessible MEROPS data base), the MTSPs protease domains share a high degree of amino acid sequence identity. The His, Asp and Ser residues necessary for activity are present in conserved motifs. The activation site, which results in the N-terminus of second chain in the two chain forms is has a conserved motif and readily can be identified (see, e.g., amino acids 801-806, SEQ ID No. 62, amino acids 406-410, SEQ ID No. 64; amino acids 186-190, SEQ ID No. 66; amino acids 161-166, SEQ ID No. 68; amino acids 255-259, SEQ ID No. 70; amino acids 190-194, SEQ ID No. 72).

The specification also describes how to identify a protease domain of the MTSPs (see, e.g., page 8):

The protease domains as provided herein are single-chain polypeptides, with an N-terminus (such as IV, VV, IL and II) generated at the cleavage site (generally have the consensus sequence R ↓VVG, R ↓VGG, R ↓ILGG, R ↓VGLL, R ↓ILGG or a variation thereof; an N-terminus of R ↓V or R ↓I, where the arrow represents the cleavage point) when the zymogen is activated. To identify a protein domain an RI should be identified, and then following amino acids compared to the above noted motif[s]. [emphasis added]

The instant specification teaches that the protease domain includes as a common structural feature a conserved catalytic triad. The art of record evidences that this is a characteristic feature. For example, Lin *et al.* teaches that membrane-type serine proteases include an invariant catalytic triad, a characteristic disulfide pattern and a proteolytic activation site in an Arg-Val-Val-Gly-Gly motif similar to the characteristic RIVGG motif in other serine proteases. (Lin *et al.*, J Biol Chem 274(26): 18231-18236 (1999), Exhibit 21). Kitamoto *et al.* teaches that the catalytic domain of MTSPs has a characteristic disulfide bond pattern (Kitamoto *et al.*, Proc Natl Acad Sci USA 91: 7588-7592 (1994), Exhibit 17). The specification teaches how to identify members of the MTSP family. For example, page 49, lines 3-10 or the specification recites:

The MTSPs are a family of transmembrane serine proteases that are found in mammals and also other species that share a number of common structural features including: a proteolytic extracellular C-terminal domain; a transmembrane domain, with a hydrophobic domain near the N-terminus; a short cytoplasmic domain; and a variable length stem region containing modular domains. The proteolytic domains share sequence homology including conserved his, asp, and ser residues necessary for catalytic activity that are present in conserved motifs.

Accordingly, the specification and the prior art sets forth specific structural and physical features that define MTSPs and their protease domains.

5. The specification provides relevant identifying characteristics of the genus

In addition to describing known and newly provided protease domains, the specification provides relevant identifying characteristics of the “genus” of serine protease domains as instantly claimed, including conserved structural and functional characteristics of an MTSP protease domain, provides a number of exemplary protease domains, and also directs those skilled in the art to exemplary art that describes common structural and functional features shared by the protease domain of MTSPs. For example, see page 26, lines 13-25, which recites:

Hence smaller portions of the protease domains, particularly the single chain domains, thereof that retain protease activity are contemplated. Such smaller versions will generally be C-terminal truncated versions of the protease domains. The protease domains vary in size and constitution, including insertions and deletions in surface loops. Such domains exhibit conserved structure, including at least one structural feature, such as the active site triad, primary specificity pocket, oxyanion hole and/or other features of serine protease domains of proteases. Thus, for purposes herein, the protease domain is a single chain portion of an MTSP, as defined herein, but is homologous in its structural features and retention of sequence of similarity or homology the protease domain of chymotrypsin or trypsin. Most significantly, the polypeptide will exhibit proteolytic activity as a single chain.

The specification teaches that included among the conserved features of MTSP protease domain polypeptides is a catalytic triad and an activation cleavage site, which defines the terminus of the protease domain polypeptides when they are isolated as single chain polypeptides.

The specification explains that beyond such conserved features, the polypeptides are tolerant of modification. The specification explains that such modifications can be effected using numerous methods known in the art. For example, at page 77, line 17 through page 78, line 11, the specification states:

A variety of modifications of the MTSP proteins and domains are contemplated herein. An MTSP-encoding nucleic acid molecule can be modified by any of numerous strategies known in the art (Sambrook *et al.*, 1990, Molecular Cloning, A Laboratory Manual, 2d ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York). The sequences can be cleaved at appropriate sites with restriction endonuclease(s), followed by further enzymatic modification if desired, isolated, and ligated in vitro. In the production of the gene encoding a domain, derivative or analog of MTSP, care should be taken to ensure that the modified gene retains the original translational reading frame, uninterrupted by translational stop signals, in the gene region where the desired activity is encoded.

Additionally, the MTSP-encoding nucleic acid molecules can be mutated in vitro or in vivo, to create and/or destroy translation, initiation, and/or termination sequences, or to create variations in coding regions and/or form new restriction endonuclease sites or destroy pre-existing ones, to facilitate further in vitro modification. Also, as described herein muteins with primary sequence alterations, such as replacements of Cys residues and elimination of glycosylation sites are contemplated. Such mutations may be effected by any technique for mutagenesis known in the art, including, but not limited to, chemical mutagenesis and in vitro site-directed mutagenesis (Hutchinson *et al.*, *J. Biol. Chem.* 253:6551-6558 (1978)), use of TAB[®] linkers (Pharmacia). In one embodiment, for example, an MTSP protein or domain thereof is modified to include a fluorescent label. In other specific embodiments, the MTSP protein is modified to have a heterofunctional reagent, such heterofunctional reagents can be used to crosslink the members of the complex.

The specification incorporates by reference and directs those skilled in the art to exemplary art that describes common structural and functional features shared by the protease domain of MTSPs. For example, Lin *et al.* (*J. Biol. Chem.* 274:18231-36 (1999), Exhibit 20) and Yan *et al.* (*J. Biol. Chem.* 274:14926-35 (1999), Exhibit 44) teach that MTSPs have highly conserved structures, including a cleavage site at the N-terminus of the protease domain, a substrate specificity pocket in the protease domain and highly conserved cysteines that participate in disulfide bonding (see, e.g., Figure 4 and page 18235 of Lin *et al.* (Exhibit 20) and Figure 2 and page 18236 of Yan *et al.* (Exhibit 44)). Other conserved elements include a conserved activation motif ((R/K)VIGG), residues Asp627, Gly-655 and Gly-665 in the substrate pocket, with Asp at the bottom of the substrate pocket, and eight conserved cysteines that form intramolecular disulfide bonds (Lin *et al.* *J Biol Chem* 274(26): 18231-18236 (1999), Exhibit 20). In addition, a correlation between retention of the catalytic triad and retention of serine protease activity was demonstrated and known in the art at the time of filing. For example, Craik *et al.* (*Science* 237: 909-913 (1987), Exhibit 10), Sprang *et al.* (*Science* 237: 905-909 (1987), Exhibit 35), Carter *et al.* (*Nature* 332: 564-568 (1988), Exhibit 8) and Bachovchin *et al.* (*Proc. Natl Acad. Sci.* 78: 7323-7326 (1981), Exhibit 5) teach that serine protease activity is retained in an MTSP by retaining the conserved structure of the catalytic triad.

The specification provides methods for identification, production, isolation, synthesis and/or purification of MTSP protease domains (see *e.g.*, working examples 1-4, which describes cloning and expression of the protease domains with the Cys replaced; Example 5 demonstrates assays for identifying inhibitors of the catalytic activity of each). The specification states, for example, that MTSP3, MTSP4 and MTSP6 are isolated from any animal, particularly a mammal, and includes but are not limited to, humans, rodents, fowl,

ruminants and other animals (see page 20, lines 21-23; page 21, lines 11-13; and page 21, lines 29-31, respectively). Alternative methods for obtaining the MTSP protein than by directly isolating the MTSP protein also are provided. These include synthesis using genomic DNA, chemically synthesizing the gene sequence from a known sequence and making cDNA to the mRNA that encodes the MTSP protein, for example, and inserting the isolated nucleic acids into an appropriate cloning vector (for example, see pages 67-79). Methods of identifying and isolating serine protease domains from MTSPs, such as MTSP1 and matriptase (also referred to as TAGD-15), corin, enteropeptidase, human airway trypsin-like protease (HAT), hepsin, TMPRS2 and TMPRSS4, were known in the art at the time of filing the application and are taught in the specification (*e.g.*, see page 4, line 20 through page 5, line 17).

In addition, the specification provides exemplary assays in which catalytic activity of the polypeptides can be tested (*e.g.*, see Examples 3 and 4). Thus, the specification describes the sequences and provides references, which are incorporated by reference, describing all of the known members of the MTSP family and the protease domains thereof, teaches how to identify an MTSP, teaches how to identify the protease domain of an MTSP if it is not known and teaches how to test the polypeptide for proteolytic activity.

The art of record and discussed previously and herein evidences that, with the information provided in the specification, the skilled artisan can recognize the protease domain of an MTSP by its requisite protease domain structure and conserved features. If necessary, one of skill in the art could test the polypeptides for catalytic activity using the assays provided in the specification or known to those of skill in art to order to identify those polypeptides that possess the requisite catalytic activity.

6. Specification describes modification of MTSP protease domains

As discussed above, a correlation between retention of the catalytic triad and retention of serine protease activity was demonstrated and known in the art at the time of filing (*e.g.*, see Craik *et al.* (Science 237: 909-913 (1987), Exhibit 10). The specification teaches additional modifications of the MTSP polypeptides such that protease activity is retained. For example, the specification explains that for each individual MTSP, the polypeptides can include about 60% amino acid sequence identity with the exemplified MTSP. Such modified polypeptides exhibit serine protease activity as single chain polypeptides. The specification provides exemplary modifications including conservative amino acid substitution (for example, see page

10, lines 3-13) and modifications of cysteine residues and/or of glycosylation sites (for example, see page 78, lines 1-7). The specification also discloses that non-natural amino acids can be introduced as a substitution or addition in the MTSP polypeptides (for example, see page 79, lines 10-21). The specification also directs those skilled in the art to exemplary art that describes common structural features shared by the transmembrane serine proteases (for example, see page 18, lines 1-15).

The specification exemplifies the replacement of a free Cys in the protease domain with another amino acid. For example, the specification states on page 10, lines 3-13 that:

Also provided are muteins of the single chain protease domains and MTSPs, particularly muteins in which the Cys residue in the protease domain that is free (i.e., does not form disulfide linkages with any other Cys residue in the protein) is substituted with another amino acid substitution, preferably with a conservative amino acid substitution or a substitution that does not eliminate the activity, and muteins in which a glycosylation site(s) is eliminated. Muteins in which other conservative amino acid substitutions in which catalytic activity is retained are also contemplated (see, e.g., Table 1, for exemplary amino acid substitutions). See, also, FIG. 4, which identifies the free Cys residues in MTSP3, MTSP4 and MTSP6.

The specification specifically describes the replacement of a free Cys in the protease domain with another amino acid. For example, Example 1, on page 161, lines 4-9, exemplifies replacing the free Cys in the protease domain with another amino acid:

To eliminate the free cysteine (at position 310 in SEQ ID No. 4) that exists when the protease domain of the MTSP3 protein is expressed or the zymogen is activated, the free cysteine at position 310 (see SEQ ID No. 3), which is Cys122 if a chymotrypsin numbering scheme is used, was replaced with a serine.

As discussed below in more detail, working examples for expression of the protease domains of MTSP3, MTSP1 and both MTSP4 are provided.

Conclusion

The claims are directed to isolated single chain protease domains of a known family of proteins, the MTSP family. The instant application provides the sequences of 17 of the known MTSP family members (directly or by incorporation by reference of references providing the sequences). The instant specification provides new members of the MTSP family and provides working examples providing the isolated protease domains thereof, where the free Cys is replaced with another amino acid. Appellant has discovered that the isolated single chain form of the protease domain of these polypeptides is active and, its use, for example, for preparing antibodies specific thereto and in diagnostic assays. Hence, the

recitation in the claims that the polypeptides consist of a protease domain from an MTSP, are single-chain polypeptides having serine protease activity and have a free Cys in the protease domain replaced with another amino acid indicates with specificity what the generic claims encompass. One skilled in the art can distinguish such a polypeptide from others and can identify species that the claims encompass. Having taught the skilled artisan that the single chain protease domain of an MTSP is active, how to identify an MTSP and its protease domain, and how to test for activity, the skilled artisan is in possession of the entire genus of single chain protease domains.

An adequate written description for a claimed genus only has to provide "relevant, identifying characteristics" of a representative number of species (MPEP §2163). It respectfully submitted that the instant specification meets this test. As noted, the specification describes all 17 known species of MTSPs and isolated protease domains (*e.g.*, see pages 9-10), as well as previously unknown species (MTSP3, MTSP4 (2 splice variants) and MTSP6), incorporates publications describing all known family members and their full length sequences, and provides relevant structural and functional features that uniquely identify and specify the claimed genus of polypeptides. The specification teaches that those of skill in the art recognize common elements among MTSPs and the protease domains of MTSPs, and teaches a number of conserved characteristics for the MTSPs and protease domains thereof, and that the sequences and locus of the protease domains are known or can be determined as taught in the application. The specification teaches that members of the MTSP family are and were known, provides additional members, teaches how to identify and isolate protease domains as single chains and how to assess activity. One of skill in the art could, if needed, readily test any of those polypeptides for catalytic activity.

Therefore, in light of Appellant's disclosure, one of skill in the art would have recognized from reading the application that Appellant provided single-chain polypeptides with the recited protease domain structure that possess serine protease activity. The combination of the disclosure of the specific chemical structures of all 17 species of MTSPs known at the time of filing and the provision and description of new species within the scope of the claims as well as teachings in the specification (and knowledge of those of skill in the art) of how to identify serine protease domains, such as based on homology as known in the art and described in the specification, and how to isolate a protease domain and also assays for testing for activity and the evidence that those of skill in the art are very familiar with the MTSP structure and

function renders it clear that one of skill in the art would recognize that Appellant had possession of the claimed polypeptides at the time of the priority date of each claim. One of skill in the art would have recognized from reading the disclosure that Appellant had possession of this genus as well as numerous species thereof. This teaching and knowledge coupled with the ability to test for species within the scope of the claims with the assays provided for in the specification and known in the art demonstrates that Appellant sufficiently described and was in possession of the polypeptides as claimed, at the effective filing date(s) of the claims.

For the reasons above, each of the dependent claims meets the written description requirement and, in addition, additional reasons for each dependent claim are described below.

Dependent Claim 11

Claim 11 depends from claim 1 and includes every limitation thereof. Claim 11 recites that the MTSP is selected from among MTSP1, MTSP3, MTSP4 and MTSP6. The specification describes MTSP1, *e.g.*, at pages 54-58. The specification describes MTSP3, *e.g.*, at pages 58-60 and Example 1 (pages 160-167). The specification describes MTSP4, *e.g.*, at pages 60-63 and Example 2 (pages 167-171). The specification describes MTSP6, *e.g.*, at pages 63-64 and Example 3 (pages 171-176). The working examples provide isolated protease domains with the free Cys residue replaced with another amino acid. Working Example 1 describes preparation and cloning and expression of the protease domain of MTSP3, Example 2 and 4, describe cloning and expression of the protease domains of MTSPs 3 and 4, and Example 3 describes cloning of MTSP6. Example 4 describes expression of the MTSP4 (both variants), MTSP3 and MTSP6 protease domains, with the replaced Cys. Example 6 describes cloning and isolated of the protease domain of MTSP1. Example 7 describes production of the protease domain of MTSP1 and purification of the protease domain.

Appellant respectfully submits that, in view of the arguments set forth above with respect to claim 1 and the teaching in the specification, which describes each of the isolated protease domains of MTSP1, MTSP3, MTSP4 (two splice variants) and MTSP6, where the free Cys is replaced with another amino acid, one of skill in the art would recognize that Appellant was in possession of the subject matter of claim 11 at its effective filing date.

Dependent Claim 20

Claim 20 depends from claim 1 and includes every limitation thereof. Claim 20 recites that a free Cys in the protease domain is replaced with a serine. For the reasons articulated above with respect to claim 1, Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of a substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid.

The specification exemplifies the replacement of a free Cys in the protease domain with serine. For example, the specification states on page 10, lines 3-13 that:

Also provided are muteins of the single chain protease domains and MTSPs, particularly muteins in which the Cys residue in the protease domain that is free (i.e., does not form disulfide linkages with any other Cys residue in the protein) is substituted with another amino acid substitution, preferably with a conservative amino acid substitution or a substitution that does not eliminate the activity, and muteins in which a glycosylation site(s) is eliminated. Muteins in which other conservative amino acid substitutions in which catalytic activity is retained are also contemplated (see, e.g., Table 1, for exemplary amino acid substitutions). See, also, FIG. 4, which identifies the free Cys residues in MTSP3, MTSP4 and MTSP6.

Table 1 of the specification identifies serine as a substitution for Cys (see page 34, line 6). The specification specifically describes the replacement of a free Cys of the protease domain with a serine in Example 1, which recites, on page 161, lines 4-9:

To eliminate the free cysteine (at position 310 in SEQ ID No. 4) that exists when the protease domain of the MTSP3 protein is expressed or the zymogen is activated, the free cysteine at position 310 (see SEQ ID No. 3), which is Cys122 if a chymotrypsin numbering scheme is used, was replaced with a serine.

Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of a substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with a serine.

Dependent Claim 34

Claim 34 depends from claim 1 and includes every limitation thereof. Claim 34 recites that the MTSP is selected from among corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4. For the reasons articulated above with respect to claim 1, Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of a substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid.

The specification specifically recites that the protease domains can be from any MTSP family member, including corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4. For example, see page 8, line 30 through page 10, line 2, which recites:

The protease domains provided herein include, but are not limited to, the single chain region having an N-terminus at the cleavage site for activation of the zymogen, through the C-terminus, or C-terminal truncated portions thereof that exhibit proteolytic activity as a single-chain polypeptide in *in vitro* proteolysis assays, of any MTSP family member, preferably from a mammal, including and most preferably human, that, for example, is expressed in tumor cells at different levels from non-tumor cells, and that is not expressed on an endothelial cell. These include, but are not limited to: MTSP1 (or matriptase), MTSP3, MTSP4 and MTSP6. Other MTSP protease domains of interest herein, particularly for use in *in vitro* drug screening proteolytic assays, include, but are not limited to: corin (accession nos. AF133845 and AB013874; see, Yan et al. (1999) J. Biol. Chem. 274:14926-14938; Tomita et al. (1998) J. Biochem. 124:784-789; Uan et al. (2000) Proc. Natl. Acad. Sci. U.S.A. 97:8525-8529; SEQ ID Nos. 61 and 62 for the human protein); enteropeptidase (also designated enterokinase; accession no. U09860 for the human protein; see, Kitamoto et al. (1995) Biochem. 27: 4562-4568; Yahagi et al. (1996) Biochem. Biophys. Res. Commun. 219:806-812; Kitamoto et al. (1994) Proc. Natl. Acad. Sci. U.S.A. 91:7588-7592; Matsushima et al. (1994) J. Biol. Chem. 269:19976-19982; see SEQ ID Nos. 63 and 64 for the human protein); human airway trypsin-like protease (HAT; accession no. AB002134; see Yamaoka et al. J. Biol. Chem. 273:11894-11901; SEQ ID Nos. 65 and 66 for the human protein); hepsin (see, accession nos. M18930, AF030065, X70900; Yamaoka et al. (1988) J Biol Chem 27: 11895-11901; Vu et al. (1997) J. Biol. Chem. 272:31315-31320; and Farley et al. (1993) Biochem. Biophys. Acta 1173:350-352; SEQ ID Nos. 67 and 68 for the human protein); TMPRSS2 (see, Accession Nos. U75329 and AF113596; Paoloni-Giacobino et al. (1997) Genomics 44:309-320; and Jacquinet et al. (2000) FEBS Lett. 468: 93-100; SEQ ID Nos. 69 and 70 for the human protein) TMPRSS4 (see, Accession No. NM 016425; Wallrapp et al. (2000) Cancer 60:2602-2606; SEQ ID Nos. 71 and 72 for the human protein); and TADG-12 (also designated MTSP6, see SEQ ID Nos. 11 and 12; see International PCT application No. WO 00/52044, which claims priority to U.S. application Ser. No. 09/261,416).

Hence, the application specifically describes the protease domain of MTSP family members corin, enteropeptidase, HAT, TMPRSS4 and TMPRSS2 and others. Appellant respectfully submits that, in view of the arguments set forth above with respect to claim 1 and the teaching in the specification, which describes the protease domain of each of corin, enteropeptidase, HAT, TMPRSS4 and TMPRSS2, one of skill in the art would recognize that Appellant was in possession of the subject matter of claim 34 at its effective filing date.

Dependent Claim 35

Claim 35 recites a conjugate that includes a) a polypeptide of claim 1, and b) a targeting agent linked to the protein directly or via a linker, wherein the conjugate has serine protease activity. For the reasons articulated above with respect to claim 1, Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of a substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid.

The specification specifically discloses conjugates of single-chain protease domains conjugated to a targeting agent, *e.g.*, at page 14, lines 19-26. The specification teaches that the conjugates can be prepared by chemical conjugation, recombinant DNA technology or combinations thereof, and provides detailed descriptions of chemical conjugation, including acid cleavable, photo-cleavable and heat sensitive linker technology and other linkers, fusion proteins, peptide linkers, conjugation to targeting agents, and adsorption, absorption and/or covalent bonding to a solid support (see *e.g.*, pages 123-131).

Appellant respectfully submits that that, in view of the arguments set forth above with respect to claim 1 and the teaching in the specification, which describes conjugates of single-chain protease domains conjugated to a targeting agent, several different types of conjugation technologies for making the conjugates and exemplary conjugates, one of skill in the art would recognize that Appellant was in possession of the subject matter of claim 35 at its effective filing date.

Dependent Claim 36

Claim 36 depends from claim 35 and recites a conjugate that includes a targeting agent that permits i) affinity isolation or purification of the conjugate; ii) attachment of the

conjugate to a surface; iii) detection of the conjugate; or iv) targeted delivery to a selected tissue or cell. For the reasons articulated above with respect to claims 1 and 35, Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of a conjugate that includes a substantially purified single-chain polypeptides consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid and a targeting agent.

The specification recites, *e.g.*, at page 14, lines 19-26 and page 123, line 30 through page 124, line 7, that the targeting agent of the conjugate permits affinity isolation or purification of the conjugate; attachment of the conjugate to a surface; detection of the conjugate; or targeted delivery to a selected tissue or cell. The specification teaches exemplary targeting agents, including tissue specific or tumor specific monoclonal antibodies, a growth factor or fragment thereof, such as FGF, EGF, PDGF, VEGF, cytokines, including chemokines, and other such agents, a protein or peptide fragment that contains a protein binding sequence, a nucleic acid binding sequence, a lipid binding sequence, a polysaccharide binding sequence, or a metal binding sequence, or a linker for attachment to a solid support (see, *e.g.*, page 124, lines 8-17) as well as linkers that allow for attachment of the conjugate to a surface (see, *e.g.*, pages 131-136). The specification also describes the construction of affinity binding pairs for isolation and/or purification of the conjugate (*e.g.*, see page 131, lines 5-37).

Appellant respectfully submits that that, in view of the arguments set forth above with respect to claims 1 and 35 and the teaching in the specification, which describes several different types of targeting agents and methods of conjugating such targeting agents to isolated protease domains, one of skill in the art would recognize that Appellant was in possession of the subject matter of claim 36 at its effective filing date.

Dependent Claim 40

Claim 40 recites a solid support comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker. For the reasons articulated above with respect to claim 1, Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of a substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a

catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid.

The specification describes solid supports and methods for immobilizing MTSP protein to solid supports (*e.g.*, see pages 131-136). The specification teaches exemplary solid supports, including supports having any required structure and geometry, such as beads, pellets, disks, capillaries, hollow fibers, needles, solid fibers, random shapes, thin films and membranes (*e.g.*, page 132, lines 26-29). The specification teaches that a plurality of MTSP protease domains, including two or more protease domains, can be attached to a solid support (*e.g.*, page 132, lines 4-8).

Appellant respectfully submits that that, in view of the arguments set forth above with respect to claim 1 and the teaching in the specification, which describes several different types of solid supports and methods of conjugating isolated protease domains to solid supports, one of skill in the art would recognize that Appellant was in possession of the subject matter of claim 40 at its effective filing date.

Dependent Claim 41

Claim 41 depends from claim 40 and recites that the polypeptides comprise an array. The specification teaches that a plurality of MTSP protease domains can be attached to a solid support (*e.g.*, see page 132, lines 4-8). The instant specification defines an array as a collection of elements containing three or more members and that, as in the case for an addressable array, the members of the array can be immobilized to discrete identifiable loci on the surface of a solid phase (*e.g.*, see page 35, lines 14-20). Hence, for these reasons and the reasons articulated above with respect to claims 1 and 40, Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of an array of substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid.

Dependent Claim 42

Claim 42 depends from claim 41 and recites that the array comprises polypeptides having different MTSP protease domains. Claim 42 as originally filed recited that the array

comprises polypeptides having different MTSP protease domains. The specification teaches that a plurality of MTSP protease domains can be attached to a solid support (*e.g.*, see page 132, lines 4-8). Appellant respectfully submits that, for these reasons and the reasons articulated above with respect to claims 1, 40 and 41, one of skill in the art would recognize that Appellant was in possession of an array of substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domains or catalytically active fragments thereof are different, have serine protease activity as a single chain and a free Cys in the protease domains is replaced with another amino acid.

Dependent Claim 113

Claim 113 recites a solid support comprising two or more polypeptides of claim 12 linked thereto either directly or via a linker. Claim 12 is not rejected under 35 U.S.C. 112, first paragraph. The Examiner states that Appellant was in possession of the isolated protease domains recited in claim 12, which is directed to the substantially purified polypeptide of claim 1, where the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12.

The specification describes solid supports and methods for immobilizing MTSP protein to solid supports (*e.g.*, see pages 131-136). The specification teaches exemplary solid supports, including supports having any required structure and geometry, such as beads, pellets, disks, capillaries, hollow fibers, needles, solid fibers, random shapes, thin films and membranes (*e.g.*, page 132, lines 26-29). The specification teaches that a plurality of MTSP protease domains, including two or more protease domains, can be attached to a solid support (*e.g.*, page 132, lines 4-8).

Appellant respectfully submits that that, because the Examiner admits that Appellant was in possession of the polypeptide of claim 12 and in view of teaching in the specification, which describes several different types of solid supports and methods of conjugating isolated protease domains to solid supports, including conjugating a plurality of isolated protease domains to a solid support, one of skill in the art would recognize that Appellant was in possession of the subject matter of claim 113 at its effective filing date.

Dependent Claim 114

Claim 114 depends from claim 113 and specifies that the polypeptides comprise an array. As discussed above, claim 113 recites a solid support that includes two or more polypeptides of claim 12. Claim 12 is not rejected under 35 U.S.C. 112, first paragraph. Thus, the Examiner agrees that Appellant was in possession of the subject matter of claim 12, which is directed to the substantially purified polypeptide of claim 1, where the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12.

The specification teaches that a plurality of MTSP protease domains can be attached to a solid support (*e.g.*, see page 132, lines 4-8). The instant specification defines an array as a collection of elements containing three or more members and that, as in the case for an addressable array, the members of the array can be immobilized to discrete identifiable loci on the surface of a solid phase (*e.g.*, see page 35, lines 14-20. Hence, for the reasons discussed above with respect to claim 1 and also because the Examiner has concluded that Appellant was in possession of the subject matter of claim 12, and the specification teaches and describes the other elements of claim 114, Appellant respectfully submits that one of skill in the art would recognize that Appellant was in possession of an array of substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid and where the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12.

Summary

Appellant respectfully submits that the rejection of claims 1, 11, 20, 34-36, 40-42, 113 and 114 under 35 U.S.C. §112, first paragraph, as allegedly containing subject matter that was not described in the specification in such a way as to reasonably convey to one skilled in the art that the inventor, at the time the application was filed, had possession of the claimed subject matter, is erroneous in law and fact and, therefore, should be reversed.

REJECTION OF CLAIMS 1, 11, 20, 34-36, 40-42, 113 AND 114 UNDER 35 U.S.C. §112, FIRST PARAGRAPH – Scope of Enablement

Claims 1, 11, 20, 34-36, 40-42, 113 and 114 are rejected under 35 U.S.C. § 112, first paragraph, because the specification allegedly fails to describe the claimed subject matter in such a way as to enable one skilled in the art to make and use the claimed subject matter commensurate in scope with these claims. The Examiner states that the specification is enabling for a polypeptide that includes amino acids 615-855 of SEQ ID NO:2, amino acids 205-437 of SEQ ID NO:4, amino acids of SEQ ID NO:6 and amino acids 217-443 of SEQ ID NO:112. The Examiner alleges that the specification does not reasonably provide enablement for a polypeptide consisting of any protease domain of any MTSP or catalytically portion thereof and concludes that the claims are drawn to polypeptides having undefined structure. The Examiner alleges that predictability of which changes in a protein's amino acid structure can be tolerated requires a knowledge of and guidance with regard to the sequence as to which amino acids, if any, are tolerant to modification and which are conserved, and detailed knowledge of how the protein's structure relates to function. It is alleged that it would require undue experimentation for one of skill in the art to make such modified polypeptides with an expectation of success because the result of such modifications is unpredictable. It is further alleged that the claimed polypeptides encompass a large number of polypeptides and that the specification does not provide sufficient guidance on the nature of the changes that can be tolerated such that the proteins retain activity. In response to Appellant's arguments in the previous Response, evidencing the extensive knowledge in the art with respect to serine proteases, the Final Office Action argues that these arguments are not persuasive because the specification allegedly does not establish which specific amino acids in the protein's sequence can be modified such that the modified polypeptide continues to have proteolytic activity. The Examiner alleges that while the art may teach the general structure of MTSP and conserved amino acid sequences, protease domains, X-ray crystal structure and other attributes, such teachings "will not reduce the burden of undue experimentation on those of ordinary skill in the art." Therefore, the Final Office Action concludes, it would require undue experimentation to produce claimed polypeptides.

This rejection respectfully is traversed. The pending claims are directed to protease domains of MTSPs, a well-characterized family of proteins; there is no doubt that this family of proteins is well known and that those of skill in the art can identify members thereof. It is the instant application that teaches that the isolated single-chain protease domain possesses

protease activity and that formation of a two-chain structure (by virtue of disulfide bonding with a Cys in the protease domain, which is free in the single chain form) is not needed. Thus the issue is not identification of an MTSP, but identification of a protease domain in an MTSP. The application clearly teaches how to identify a protease domain and how to replace the now free Cys that would have participated in forming a two chain structure. There are no issues regarding undue experimentation to isolate MTSPs.

The specification teaches identification, preparation and isolation of protease domains and those of skill in the art, in view of the application, readily can identify and isolate a protease domain from any MTSP. As discussed above, with respect to the written description rejection, the claims are directed to isolated single chain protease domains. The specification teaches that those of skill in the art can identify protease domains and also teaches how to identify protease domains. One of skill in the art, in light of the specification, could prepare an isolated single chain protease domain, as claimed, for any MTSP and replace the now-free Cys with another amino acid. Hence there is no reason to limit the claims to particular species of the family, when one of skill in the art, in light of the disclosure, can identify all members of the genus.

A. LEGAL STANDARDS - 35 U.S.C. §112, FIRST PARAGRAPH – ENABLEMENT

The inquiry with respect to scope of enablement under 35 U.S.C. § 112, first paragraph, is whether it would require undue experimentation to make and use the subject matter as claimed. A considerable amount of experimentation is permissible, particularly if it is routine experimentation. The amount of experimentation that is permissible depends upon a number of factors, which include: the quantity of experimentation necessary, the amount of direction or guidance presented, the presence or absence of working examples, the nature of the invention, the state of the prior art, the relative skill of those in the art, the predictability of the art, and the breadth of the claims (i.e., the "*Wands* factors"). *In re Wands*, 8 USPQ2d 1400 (Fed. Cir. 1988).

The starting point in an evaluation of whether the enablement requirement is satisfied is an analysis of each claim to determine its scope. The focus of the inquiry is whether everything within the scope of the claim is enabled. As concerns the breadth of a claim relevant to enablement, the only concern should be whether the scope of enablement provided to one skilled in the art by the disclosure is commensurate with the scope of protection sought by the claims. *In re Moore*, 439 F.2d 1232, 169 USPQ 236 (CCPA 1971). Once the scope of the claims is

addressed, a determination must be made as to whether one skilled in the art is enabled to make and use the entire scope of the claimed invention without undue experimentation.

It is incumbent upon the Examiner to first establish a *prima facie* case of non-enablement. *In re Marzocchi*, 439 F.2d 220, 223, 169 USPQ 367, 369-70 (CCPA 1971). The requirements of 35 USC §112, first paragraph, can be fulfilled by the use of illustrative examples or by broad terminology. *In re Anderson*, 176 USPQ 331, 333 (CCPA 1973):

... we do not regard section 112, first paragraph, as requiring a specific example of everything within the scope of a broad claim ... What the Patent Office is here apparently attempting is to limit all claims to the specific examples, not withstanding the disclosure of a broader invention. This it may not do.

In re Grimme, 274 F.2d 949, 952 (CCPA 1960) :

It is manifestly impracticable for an applicant who discloses a generic invention to give an example of every species falling within it, or even to name every such species. It is sufficient if the disclosure teaches those skilled in the art what the invention is and how to practice it.

This clause does not require "a specific example of everything *within the scope* of a broad claim." *In re Anderson*, 176 USPQ 331, at 333 (CCPA 1973), emphasis in original. Rather, the requirements of § 112, first paragraph "can be fulfilled by the use of illustrative examples or by broad terminology." *In re Marzocchi et al.*, 469 USPQ 367 (CCPA 1971)(emphasis added).

The law is clear that patent documents need not include subject matter that is known in the field of the invention and is in the prior art, for patents are written for persons experienced in the field of the invention. See *Vivid Technologies, Inc. v. American Science and Engineering, Inc.*, 200 F.3d 795, 804, 53 USPQ2d 1289, 1295 (Fed. Cir. 1999) ("patents are written by and for skilled artisans"). To hold otherwise would require every patent document to include a technical treatise for the unskilled reader. Although an accommodation to the "common experience" of lay persons may be feasible, it is an unnecessary burden for inventors and has long been rejected as a requirement of patent disclosures. See *Atmel Corp.*, 198 F.3d at 1382, 53 USPQ2d at 1230 (Fed. Cir. 1999) ("The specification would be of enormous and unnecessary length if one had to literally reinvent and describe the wheel."); *W.L. Gore & Assoc., Inc. v. Garlock, Inc.*, 721 F.2d 1540, 1556, 220 USPQ 303, 315 (Fed. Cir. 1983) ("Patents are written to enable those skilled in the art to practice the invention, not the public.")

The test of enablement is whether one skilled in the art can make and use what is claimed based upon the disclosure in the application and information known to those of skill in the art without undue experimentation. *United States v. Telectronics, Inc.*, 8 USPQ2d 1217

(Fed. Cir. 1988). A certain amount of experimentation is permissible as long as it is not undue. *Atlas Powder Co. v. E.I. DuPont de Nemours*, 750 F.2d 1569, 224 USPQ 409 (1984). This requirement can be satisfied by providing sufficient disclosure, either through illustrative examples or terminology, to teach one of skill in the art how to make and how to use the claimed subject matter without undue experimentation. *In re Anderson*, 176 USPQ 331, at 333 (CCPA 1973). The "invention" referred to in the enablement requirement of section 112 is the claimed subject matter. *Lindemann Maschinen-fabrik v. American Hoist and Derrick Co.*, 730 F.2d 1452, 1463, 221 USPQ 481, 489 (Fed. Cir. 1984).

As a matter of Patent Office practice, then, a specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented *must* be taken as in compliance with the enabling requirement of the first paragraph of § 112 *unless* there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support. Assuming that sufficient reason for such doubt does exist, a rejection for failure to teach how to make and/or use will be proper on that basis; such a rejection can be overcome by suitable proofs indicating that the teaching contained in the specification is truly enabling. . . it is incumbent upon the Patent Office, whenever a rejection on this basis is made, to explain why it doubts the truth or accuracy of any statement in a supporting disclosure and to back up assertions of its own with evidence or reasoning which is inconsistent with the contested statement.

Id. (emphasis in original); *See also Fiers v. Revel*, 984 F.2d 1164, 1171-72, 25 USPQ2d 1601, 1607 (Fed. Cir. 1993); *Gould v. Mossinghoff*, 229 USPQ 1, 13 (D.D.C. 1985), *aff'd in part, vacated in part, and remanded sub nom. Gould v. Quigg*, 822 F.2d 1074, 3 USPQ2d 1302 ("there is no requirement in 35 U.S.C. § 112 or anywhere else in patent law that a specification convince persons skilled in the art that the assertions in the specification are correct"). A patent application need not teach, and preferably omits, what is well known in the art. *Spectra-Physics, Inc. v. Coherent, Inc.*, 3 USPQ2d 1737 (Fed. Cir. 1987).

PTO GUIDELINES

The PTO has promulgated guidelines, which incorporate the above-noted law, for examining chemical/biotechnical applications with respect to 35 U.S.C. § 112, first paragraph, enablement. As set forth in the guidelines, the standard for determining whether the specification meets the enablement requirement is whether it enables any person skilled in the art to make and use the claimed invention without undue experimentation. *In re Wands*, 858 F.2d 731, 737, 8 USPQ2d 1400 (Fed. Cir. 1988). In determining whether any experimentation is "undue," consideration must be given to the above-noted factors.

As indicated in the published guidelines, it is improper to conclude that a disclosure is not enabling based on an analysis of only one of the above factors while ignoring one or more of the others. The analysis must consider all the evidence related to each of the factors, and any conclusion of non-enablement must be based on the evidence as a whole. *Id.* 8 USPQ2d at 1404 & 1407.

B. THE REJECTION OF CLAIMS 1, 11, 20, 34-36, 40-42, 113 AND 114 UNDER 35 U.S.C. §112, FIRST PARAGRAPH SHOULD BE REVERSED BECAUSE THE SPECIFICATION MEETS THE WRITTEN DESCRIPTION REQUIREMENT WITH RESPECT TO ENABLEMENT

APPLICATION OF THE FACTORS ENUMERATED IN *IN RE WANDS*

Claim 1

It respectfully is submitted that analysis of enablement requires consideration of all of the “*Wands* Factors” and that focusing on one or two of the factors is a misapplication of the law. Appellant has discussed application of the “*Wands* Factors” in the previous responses. It would not require undue experimentation to isolate single-chain protease domains from any MTSP polypeptide. Further, it would not require undue experimentation to make modifications thereto. The Examiner admits that enzyme isolation techniques and recombinant and mutagenesis techniques are known in the art, and that it is routine in the art to screen for substitutions or modifications, including multiple substitutions and multiple modifications as encompassed by the instant claims (see Final Office Action, Exhibit 2, page 11). As discussed in detail below, and previously, a consideration of the factors enumerated in *In re Wands* demonstrates that the application teaches how to make and use the subject matter as claimed without undue experimentation.

i. Breadth of the Claims

Claim 1 is directed to an isolated substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, wherein the protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid. Claims 11, 20, 34-36, 40-42, 113 and 114 ultimately depend from claim 1 and recite additional features and specific family members. Claim 11 is directed to the substantially purified polypeptide of claim 1, and specifies that the MTSP is selected from among MTSP1, MTSP3, MTSP4 and MTSP6.

Claim 20 recites that a free Cys in the protease domain is replaced with a serine. Claim 34 recites particular polypeptides within the scope of claim 1. Claims 35 and 36 are

directed to conjugates including a polypeptide of claim 1 and a targeting agent linked to the protein directly or via a linker. Claims 40-42 are directed to a solid support including two or more polypeptides of claim 1 linked thereto either directly or via a linker. Claims 113 and 114 are directed to a solid support including two or more polypeptides of claim 12 linked thereto either directly or via a linker.

Hence the claims include as an element an isolated protease domain of a member of the MTSP family in which a fee Cys is replaced with another amino acid. The specification, as noted, describes all MTSP family members known at the time of filing and provides four new members of the family and methods for identifying other members of the MTSP family. Thus, the claims are of the same scope as the disclosure in the application.

ii. Level of Skill

The level of skill in this art is recognized to be high (see, *e.g.*, *Ex parte Forman*, 230 USPQ 546 (Bd. Pat. App. & Int'f 1986)). The numerous articles and patents made of record in this application address a highly skilled audience and further evidence the high level of skill in this art.

iii. Teachings of the Specification

As discussed above and previously, the specification teaches that MTSP polypeptides constitute a recognized well known and well characterized family of serine proteases. For example, page 18, lines 1-23 of the specification recites:

As used herein, "transmembrane serine protease (MTSP)" refers to a family of transmembrane serine proteases that share common structural features as described herein (see, also Hooper et al. (2001) J. Biol. Chem. 276:857-860). Thus, reference, for example, to "MTSP" encompasses all proteins encoded by the MTSP gene family, including but are not limited to: MTSP1, MTSP3, MTSP4 and MTSP6, or an equivalent molecule obtained from any other source or that has been prepared synthetically or that exhibits the same activity. Other MTSPs include, but are not limited to, corin, enteropeptidase, human airway trypsin-like protease (HAT), MTSP1, TMPRSS2, and TMPRSS4. Sequences of encoding nucleic molecules and the encoded amino acid sequences of exemplary MTSPs and/or domains thereof are set forth in SEQ ID Nos. 1-12, 49, 50 and 61-72. The term also encompasses MTSPs with conservative amino acid substitutions that do not substantially alter activity of each member, and also encompasses splice variants thereof. Suitable conservative substitutions of amino acids are known to those of skill in this art and may be made generally without altering the biological activity of the resulting molecule. Of particular interest are MTSPs of mammalian, including human, origin. Those of skill in this art recognize that, in general, single amino acid substitutions in non-essential regions of a polypeptide do not substantially alter biological activity (see, *e.g.*, Watson et al. Molecular Biology of the Gene, 4th Edition, 1987, The Benjamin/Cummings Pub. Co., p.224).

The specification teaches that a protease domain from an MTSP polypeptide is active as a single-chain polypeptide. Additionally, smaller fragments of the protease domain also are active as single-chain polypeptides (page 18, line 24-page 19, line 2):

As used herein, a "protease domain of an MTSP" refers to the protease domain of MTSP that is located within the extracellular domain of a MTSP and exhibits serine proteolytic activity. It includes at least the smallest fragment thereof that acts catalytically as a single chain form. Hence it is at least the minimal portion of the extracellular domain that exhibits proteolytic activity as assessed by standard assays *in vitro* assays. Those of skill in this art recognize that such protease domain is the portion of the protease that is structurally equivalent to the trypsin or chymotrypsin fold.

The specification further teaches that MTSP protease domains can vary in sequence but that these proteins retain a conserved structure as well as sequence identity to identified MTSP proteins exemplified in the application. For example, see page 19, lines 3-24, which recites:

Exemplary MTSP proteins, with the protease domains indicated, are illustrated in Figures 1-3. Smaller portions thereof that retain protease activity are contemplated. The protease domains vary in size and constitution, including insertions and deletions in surface loops. They retain conserved structure, including at least one of the active site triad, primary specificity pocket, oxyanion hole and/or other features of serine protease domains of proteases. Thus, for purposes herein, the protease domain is a portion of a MTSP, as defined herein, and is homologous to a domain of other MTSPs, such as corin, enteropeptidase, human airway trypsin-like protease (HAT), MTSP1, TMPRSS2, and TMPRSS4, which have been previously identified; it was not recognized, however, that an isolated single chain form of the protease domain could function proteolytically in *in vitro* assays. As with the larger class of enzymes of the chymotrypsin (S1) fold (see, e.g., Internet accessible MEROPS data base), the MTSPs protease domains share a high degree of amino acid sequence identity. The His, Asp and Ser residues necessary for activity are present in conserved motifs. The activation site, which results in the N-terminus of the second chain in the two chain forms is has a conserved motif and readily can be identified (see, e.g., amino acids 801-806, SEQ ID No. 62, amino acids 406-410, SEQ ID No. 64; amino acids 186-190, SEQ ID No. 66; amino acids 161-166, SEQ ID No. 68; amino acids 255-259, SEQ ID No. 70; amino acids 190-194, SEQ ID No. 72).

The application describes the full length sequence and protease domain of all species of MTSP family members known at the time of filing, including MTSP1, HAT, corin, enteropeptidase, TMPRSS4 and TMPRSS2. The specification also identifies four new family members.

As discussed above, identification of the protease domain from an MTSP region merely requires identification of the activation cleavage site, as is outlined in the specification, discussed above and known in the art. The locus of the protease domain in the known MTSP family members is known, and the instant application provides protease domains from the known family members, either directly or by incorporation of reference.

Furthermore, notwithstanding that the specification provides and describes the protease domain of all members of the family known at the time of filing, plus the four additional family members, a comparison of sequence identity among family members (see, e.g., Figure 4 of the

application) reveals that the protease domains share conserved sequences, including the catalytic triad of His, Asp and Ser residues and their surrounding conserved motifs. Additionally, the specification demonstrates that MTSP protease domains can have a reasonable amount of sequence variation and yet retain serine protease activity. MTSP1, MTSP3, MTSP4 and MTSP6 protease domains share about 40% sequence identity with each other. The specification teaches that each of these protease domains is an example of an MTSP protease domain that has activity in the single chain form.

The specification also teaches additional modifications. For example, see page 26, lines 13-25, which recites:

Hence smaller portions of the protease domains, particularly the single chain domains, thereof that retain protease activity are contemplated. Such smaller versions will generally be C-terminal truncated versions of the protease domains. The protease domains vary in size and constitution, including insertions and deletions in surface loops. Such domains exhibit conserved structure, including at least one structural feature, such as the active site triad, primary specificity pocket, oxyanion hole and/or other features of serine protease domains of proteases. Thus, for purposes herein, the protease domain is a single chain portion of an MTSP, as defined herein, but is homologous in its structural features and retention of sequence of similarity or homology the protease domain of chymotrypsin or trypsin. Most significantly, the polypeptide will exhibit proteolytic activity as a single chain.

The specification teaches that included in the conserved features of MTSP protease domain polypeptides is a catalytic triad as well as the activation cleavage site, which defines the terminus of the protease domain polypeptides when they are isolated as single chain polypeptides.

The specification explains that beyond such conserved features the polypeptides are tolerant of modification. The specification explains that such modifications can be effected using numerous methods known in the art. For example, at page 77, line 17 through page 78, line 11, the specification states:

A variety of modifications of the MTSP proteins and domains are contemplated herein. An MTSP-encoding nucleic acid molecule can be modified by any of numerous strategies known in the art (Sambrook et al., 1990, Molecular Cloning, A Laboratory Manual, 2d ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York). The sequences can be cleaved at appropriate sites with restriction endonuclease(s), followed by further enzymatic modification if desired, isolated, and ligated in vitro. In the production of the gene encoding a domain, derivative or analog of MTSP, care should be taken to ensure that the modified gene retains the original translational reading frame, uninterrupted by translational stop signals, in the gene region where the desired activity is encoded.

Additionally, the MTSP-encoding nucleic acid molecules can be mutated in vitro or in vivo, to create and/or destroy translation, initiation, and/or termination

sequences, or to create variations in coding regions and/or form new restriction endonuclease sites or destroy pre-existing ones, to facilitate further in vitro modification. Also, as described herein muteins with primary sequence alterations, such as replacements of Cys residues and elimination of glycosylation sites are contemplated. Such mutations may be effected by any technique for mutagenesis known in the art, including, but not limited to, chemical mutagenesis and in vitro site-directed mutagenesis (Hutchinson *et al.*, J. Biol. Chem. 253:6551-6558 (1978)), use of TAB[®] linkers (Pharmacia). In one embodiment, for example, an MTSP protein or domain thereof is modified to include a fluorescent label. In other specific embodiments, the MTSP protein is modified to have a heterofunctional reagent, such heterofunctional reagents can be used to crosslink the members of the complex.

The specification exemplifies variation in MTSP sequences. For example the specification provides exemplary MTSP1, MTSP3, MTSP4 and MTSP6 sequences, including the sequences of the isolated protease domains. The specification also provides sequences of other family members, and, as discussed above, how to identify the protease domain based on the consensus sequence thereof, which is conserved among serine proteases. The specification explains that MTSP1 and MTSP3 amino acid sequences have about 43% identity with each other (for example, see page 162, lines 1-2). The specification also discloses that MTSP1 and MTSP4 have about 37% amino acid sequence identity (for example, see page 167, lines 25-29). The specification also teaches that MTSP4 and MTSP6 share about 60% amino acid sequence identity (for example, see page 172, lines 4-9). The specification teaches that each of the protease domains of these MTSP family members is active as single chain that contains only the protease domain or a smaller catalytically active portion of the protease domain (see, for example at page 20, lines 1-6). Hence, the specification teaches that MTSP protease domains that retain the conserved catalytic triad are tolerant of sequence modification yet retain activity, and demonstrates that exemplary polypeptides that retain the catalytic triad and that have about 40%-60% and greater sequence identity are active as single chain polypeptides.

Notwithstanding differences among the sequences of the family members, the specification teaches and provides sequences of most of the family members, refers to publications that describe other family members, teaches how to identify a protease domain. As discussed above, the instant claims are not directed to discovery of MTSPs as a family, but the discovery that the isolated protease domain has activity as a single-chain isolated polypeptide. Once one of skill in the art has an MTSP of any type or sequence, one of skill in

the art, based on the teachings in this specification, isolate the single chain protease domain thereof. The specification clearly provides guidance for doing so.

The specification teaches a modifications of the MTSP polypeptides. For example, the specification provides exemplary modifications including conservative amino acid substitution (for example, see page 10, lines 3-13) and modifications of cysteine residues and/or of glycosylation sites (for example, see page 78, lines 1-7). The specification also discloses that non-natural amino acids can be introduced as a substitution or addition in the MTSP polypeptides (for example, see page 79, lines 10-21).

More significantly, the pending claims are directed, not to full-length MTSPs, but to isolated single-chain protease domains, where the free Cys is replaced with another amino acid that have serine protease activity. One of skill in the art, with an MTSP polypeptide in hand, could readily identify and isolate the protease domain of any MTSP as claimed and replace a free Cys with another amino acid residue.

iv. Knowledge of those of skill in the art

As discussed above, at the time of filing of the application and before, those of skill in the art were very familiar with serine proteases generally, and with the MTSP family in particular. The MTSP family was known as was the locus of the protease domain in members of the MTSP family. What was absent was any understanding or recognition that an isolated single chain protease domain would have activity; hence, such was never isolated. In view of the instant application teaching that such protease domains have activity as single chain polypeptides, the skilled artisan can readily isolate any protease domain of an MTSP as a single chain and if necessary test the isolated protease domain for the requisite activity. Nothing more need be known regarding the requisites for activity.

Notwithstanding this, there was a large body of literature directed to serine proteases and there was general understanding of their structures and requisites for activity (see for example, Hooper *et al.*, J. Biol. Chem. 276: 857-860 (2001), Exhibit 15; Nienaber *et al.*, J. Biol. Chem. 275: 7239-7248 (2000), Exhibit 24; Sommerhoff *et al.*, Proc. Natl. Acad. Sci. USA 96: 10984-10991 (1999), Exhibit 34; Lu *et al.*, J. Mol. Biol. 292: 361-373 (1999), Exhibit 21; Xu *et al.*, J. Biol. Chem. 275: 378-385 (2000), Exhibit 41; Lin *et al.*, J. Biol. Chem. 274: 18231-18236 (1999), Exhibit 20; and Bryan, Biochem. Biophys. Acta 1543: 200-203 (2000), Exhibit 7). These references detail the existing crystal structures, structural comparisons and structural similarities of MTSPs.

This extensive knowledge also is evidenced, for example, in the application as filed and in the literature made of record in the submitted Information Disclosure Statements. As noted in the application, the MTSP protease family was known (for example, see pages 4-5). Serine proteases are a family that can be distinguished from many other types of proteins and enzymes because they have highly conserved structures (see e.g., Lin *et al.*, J. Biol. Chem. 274: 18231-18236 (1999), Exhibit 20 and Yan *et al.*, J. Biol. Chem. 274: 14926-14935 (1999), Exhibit 44). Moreover, it was known at the time of filing that there is a known correlation between retention of the catalytic triad and retention of serine protease activity. Hence, available to one of skill in the art was the knowledge that serine protease activity could be retained in a serine protease by retaining the conserved structure of the catalytic triad (see for example, Carter *et al.*, Nature 332: 564-368 (1988), Exhibit 8, Sprang *et al.*, Science 237: 905-909 (1987), Exhibit 35, Craik *et al.*, Science 237: 909-913 (1987), Exhibit 10 and Bachovchin *et al.*, Proc. Natl Acad. Sci. 78: 7323-7326 (1981), Exhibit 5). In addition, other features were identified at the time of filing and before as highly conserved features in serine proteases including a cleavage site at the N-terminus of the protease domain, a substrate specificity pocket in the protease domain and conserved cysteines that participate in disulfide bonding (see for example, Figure 4 and page 18235 of Lin *et al.* (Exhibit 20) and Figure 2 and page 18236 of Yan *et al.*, Exhibit 44). Thus, the requisites for retention of serine protease activity are well known and characterized and were available at the effective filing date of the claimed subject matter. Hence, a wide variety of structural information on serine proteases was well-known in the art.

Furthermore, the instant claims only require identification of the protease domain of an MTSP, and its isolation as a single chain polypeptide. The specification includes and describes the protease domains of all MTSP family members known at the time of filing the application. Based on the teachings of the specification and known in the art, those of skill in the art can readily identify the protease domain region in an MTSP using, e.g., the catalytic triad, the cleavage site at the N-terminus of the protease domain and conserved cysteines that participate in disulfide bonding as markers, and, if necessary test it for protease activity. Dawson *et al.* (U.S. Pat. No. 5,645,833 (1997), Exhibit 11) teaches that the serine protease domain can be recognized by its homology with other serine proteases (col. 6, lines 29-32).

The methods and guidance for comparing amino acid sequences to generate and confirm sequences with sequence identity to an MTSP polypeptide sequence such as SEQ ID NOS: 2, 4, 6 and 12 was available and routine in the art at the time of filing the instant

application. As described in the instant specification, computer algorithms such as the "FAST A" program, using for example, the default parameters as in Pearson *et al.*, Proc. Natl. Acad. Sci. USA 85: 2444 (1988), Exhibit 28, were available. Other programs were available (see Devereux, J., *et al.*, Nucleic Acids Research 12(I):387 (1984), Exhibit 12). In addition, methods for generating nucleotide and protein sequence variation were widely available in the art. Thus, one of skill in the art could use such programs with a serine protease sequence, for example, to align the sequence and identify the structural features of importance for retention of activity and use the methods for generating sequence variation to make protein variants.

Methods for assaying protease activity including protease specificity, level of activity and response to inhibitors was well known in the art (see, for example, Lu *et al.*, J. Mol. Biol. 292: 361-373 (1999) (Exhibit 21) and Xu *et al.*, J. Biol. Chem. 275: 378-385 (2000) (Exhibit 41)). Methods for high throughput assays and detection also were widely available (e.g., see generally, Silverman *et al.*, Curr. Opin. Chem. Biol., 2:397-403 (1998) (Exhibit 32) and Sittampalam *et al.*, Curr. Opin. Chem. Biol., 1:384-91 (1997) (Exhibit 33). Hence, the amount of knowledge of those of skill in the art was extensive and the requisite structural and functional features required for protease activity was well known.

The Examiner states that the specific amino acid positions within a protein's sequence where amino acid modification can be made with a reasonable expectation of success in obtaining the desired activity are limited in any protein and the result of such modifications is unpredictable. Appellant respectfully disagrees in the case of the family of MTSPs. The application and the art made of record establish that MTSPs are well known in the art and the structural requirements for activity are known and that the instantly claimed polypeptides share sequence homology with the chymotrypsin/trypsin family for which tertiary structures are known. For example, it was known in the art that serine protease activity could be retained in an MTSP by retaining the conserved structure of the catalytic triad (see e.g., Craik *et al.*, Science 237: 909-13 (1987), Exhibit 1 and Carter *et al.*, Nature 332: 564-568 (1988), Exhibit 8). Other highly conserved features in serine proteases also were known to the skilled artisan. These include a cleavage site at the N-terminus of the protease domain, a substrate specificity pocket in the protease domain and conserved cysteines that participate in disulfide bonding (see, e.g., Figure 4 and page 18235 of Lin *et al.* (Exhibit 20) and Figure 2 and page 18236 of Yan *et al.* (Exhibit 44). The specification also provides exemplary assays for testing

catalytic activity of the polypeptides using routine experimental analysis techniques and also provides descriptions of how to assess percentage identity and teaches that these techniques were well known in the art. The specification also teaches conserved characteristics among MTSPs. Furthermore, the MTSPs are a known family of serine proteases, and the protease domain of any member can be readily identified using methods and techniques known in the art and/or described in the specification. The serine proteases were among the first enzymes to be studied extensively (Perona & Craik, Protein Science 4: 337-360 (1995), Exhibit 30).

Furthermore, the instant claims are directed to the single-chain protease domain or active portion thereof, where protease domain is modified to replace a free Cys with another amino acid (for example to prevent aggregation by virtue of interaction among the free Cys residues). The claims on appeal are not new MTSPs per se, but to the protease domains of MTSPs.

The Examiner states that recombinant and mutagenesis techniques and enzyme isolation techniques are known and that it is routine to screen for multiple substitutions or multiple modifications as encompassed by the instant claims (see Final Office Action, Exhibit 1, page 11). Thus, routine techniques can be used to identify or synthesize modified MTSP serine protease domains. If needed, one of skill in the art can test polypeptides for catalytic activity by routine experimentation using the assays provided in the specification or known to those of skill in art.

v. Working Examples

The application provides working examples that demonstrate each of the features of the claimed polypeptides. For instance, the Examples provide detailed guidance for identifying and isolating MTSP protease domains. Example 1 describes the cloning of the full-length and the protease domain of MTSP3 and replacement of the free Cys in the isolated protease domain with another amino acid. Example 1 also describes expression of the MTSP3 protease domain with replaced Cys. Example 1 also describes the use nucleic acid encoding the probe to assess tissue-specific and tumor-specific expression of the MTSP3.

Example 2 describes the identification and cloning of two MTSP4 polypeptides, MTSP4-S and MTSP4-L. Example 2 describes cloning of the full-length polypeptides and also the protease domains thereof, and also describes uses of the clones to obtain gene expression profiles. Example 3 describes the identification and cloning of an MTSP6 polypeptide and protease domain thereof, and also gene expression profiles. Example 4 describes expression of

the MTSP4 (both variants), MTSP3 and MTSP6 protease domains, with the replaced Cys. Example 6 describes cloning and isolated of the protease domain of MTSP1. Example 7 describes production of the protease domain of MTSP1 and purification of the protease domain. In each case, an MTSP polypeptide sequence is identified that includes a protease domain with a cleavage site and a catalytic triad (see, e.g., Figure 4). As noted, for example, in Example 1, identification of MTSP3 as a serine protease required only 43% sequence identity. Similarly, Example 2 demonstrates that 37% sequence identity with MTSP1 was sufficient to identify MTSP4.

The Examples demonstrate additional features of the claimed polypeptides. For example, the examples demonstrate production and expression of MTSP protease domains, where the free Cys is replaced with another amino acid. The working examples further demonstrate that the MTSP polypeptides, sharing, for example, 37-43% sequence identity, are active as a single chain protease domains.

The Examples demonstrate expression of single chain protease domains. Examples 4 and 5 describe additional expression of MTSP3, MTSP4 and MTSP6 using *Pichia pastoris*. Examples 6 and 7 provide a detailed description of the cloning, expression and purification of an MTSP1 single chain protease domain. Example 8 provides detailed serine protease assays for MTSP1. Additionally, the examples demonstrate replacement of the free Cys. For example, Example 1 demonstrates that replacing the cysteine to serine does not substantially alter serine protease activity. The examples demonstrate identification of a variety of MTSPs, sharing 37-43% sequence identity, and the expression of the protease domains thereof, where the Cys is replaced with another amino acid.

vi. Predictability

The predictability at issue is whether one of skill in the art could isolate protease domains from MTSP family members and variants thereof. The issue is not whether the claims encompass variant MTSPs, but whether one of skill in the art in possession of an MTSP could prepare an isolated protease domain in which a free Cys is replaced with another amino acid. Predictability goes to reproducibility. Issues regarding modification of MTSPs and requisites therefore are irrelevant. Appellant respectfully submits that one of skill in the art, given the instant disclosure, could predictably make such polypeptides, because the MTSP family is well known and characterized and the sequences of exemplary new family members, as well as all known members, are provided in the application. One of skill in the

art readily make minor amino acid variation using routine techniques, and, if needed, test such polypeptide variants for serine protease activity. The working example demonstrate repeating this with 5 different polypeptides (MTSP1, MTSP3, MTSP4-S, MTSP4-L and MTSP6). There is no doubt that isolation of a protease domain from an MTSP is reproducible and, thus, predictable. There is no doubt that one of skill in the art could prepare an isolated protease domain as claimed using techniques routinely practiced in this art.

In contrast to the allegations of “unpredictability” set forth in the Final Office Action, the specification and the knowledge in the art evidence many factors of predictability with respect to MTSP polypeptide variants. First, the specification identifies all known MTSP family members, including the sequences thereof (in the sequence listing and/or by incorporation by reference of others) and also provides new family members. These are defined chemical structures from which one of skill in the art is given a reference point. As explained above, included among exemplary polypeptides are MTSP1, MTSP3, MTSP4-S, MTSP4-L, MTSP6, HAT, corin, enteropeptidase, TMPRSS4 and TMPRSS2. The specification demonstrates that these MTSP polypeptides, as well as all family members, share conserved features including a protease domain with a catalytic triad and N-terminal activation cleavage site. Furthermore, the specification teaches isolation of the protease domains as single chains and demonstrates that they possess proteolytic activity. As discussed above, the specification provides detailed guidance for identifying a protease domain of any MTSP family member.

Second, the specification delineates structural and functional features of the protein. These features identify key regions and residues that one of skill in the art would know to conserve in order to retain serine protease activity. These features also provide reference points for alignments with other known serine proteases. These features also allow one of skill in the art to make further structure-function correlations, again providing predictable correlations of regions and residues to conserve or change. As evidenced by the references cited in the specification and in the Information Disclosure Statements of record in this application and provided herein, a large body of knowledge pertaining to structure-function relationships of serine proteases was known in the art. In addition, the specification provides exemplary assays to assess serine protease activity, including a variety of substrates, for MTSP activity. One of skill in the art can readily and routinely test any MTSP family

member protease domain or a variant thereof for serine protease activity as a single chain protease.

As taught in the specification as well as evidenced by the art of record, maintenance of the catalytic triad is sufficient to retain serine protease activity (e.g., see Carter *et al.* (Nature 332: 564-568 (1988), Exhibit 8 and Craik *et al.* (Science 237: 909-913 (1987), Exhibit 10)). Therefore, one of skill in the art could make and generate MTSP family member protease domains from any MTSP known to one of skill in the art or identify protease domains in new MTSP family members. In the unlikely event that it was needed, protease activity could easily and routinely be confirmed using the assays provided in the application and known in the art. The routine manipulations to identify and isolate an MTSP protease domain as a single chain are known in the art.

The experimentation necessary to isolate and use protease domains of MTSP polypeptides, as described above, is commonly practiced in this art and routine. "Enablement is not precluded by the necessity for some experimentation such as routine screening. Experimentation needed to practice the invention must not be undue experimentation. 'The key word is *undue*, not experimentation.' " *In re Wands*, 858 F.2d at 737-38 (quoting *In re Angstadt*, 537 F.2d at 504; emphasis added; additional internal citations omitted). The Examiner admits that enzyme isolation techniques and recombinant and mutagenesis techniques are known and that it is routine to screen for multiple substitutions or multiple modifications as encompassed by the instant claims (see Final Office Action, Exhibit 2, page 11). The art related to serine proteases also demonstrates that such experimentation is not undue. For example, Pearson *et al.* (Cabios Invited Review 13(4): 325-332 (1997) (Exhibit 29)) explains that serine proteases share a conserved catalytic site, the catalytic triad and have several diagnostic motifs throughout the protein including a conserved protein fold and anti-parallel β barrel structures that contribute to the function of the protease. Pearson *et al.* states that one could recognize proteins that have protease activity based on these conserved structures. Hence, generation of variants with serine protease activity is routine because one of skill in the art can use such conserved features as a guide for designing the location of variations to maintain these features. In addition, Cheah *et al.* (J. Biol. Chem. 265: 7180-7187 (1990), Exhibit 9) provides a demonstration of the predictability of generating variants of serine proteases based on an exemplary sequence. Cheah *et al.* uses known structural and functional information about trypsin-like serine proteases to obtain mutations in a rhinovirus

3C protease with predicted functional phenotypes. Thus, the art available at the time of filing, and before, demonstrates that one of skill in the art could make variants of a serine protease in a predictable manner. Therefore, one of skill in the art could make protease domains as single chains from an MTSP family member and also generate variants of MTSP polypeptides, using routine biotechnology techniques. Activity of the single chain protease domains and variants thereof could easily and routinely be confirmed using the assays provided in the application and known in the art. The routine manipulations to generate an MTSP single chain protease domain are not unpredictable.

As discussed above, the issue is not whether the claims encompass variant MTSPs, but whether one of skill in the art in possession of an MTSP could prepare an isolated protease domain in which a free Cys is replaced with a another amino acid. The instant application identifies MTSP polypeptides and exemplifies that isolated serine protease domains possess serine protease activity as a single chain. Such demonstration of single chain activity had not been demonstrated before the instant application. The application provides adequate description to demonstrate that a common feature among the MTSP family members is the activity of a single chain form that includes the protease domain or catalytically active portions thereof in the absence of other MTSP portions. The application provides exemplary MTSP's that share about 40% sequence identity and possess such features. As discussed, the working examples, demonstrate reproducibility, producing 5 different protease domains. Therefore, the specification demonstrates that by following the teachings of the application, one of skill in the art can predictably identify, make and use substantially purified polypeptides consisting of an MTSP protease domain or catalytically active fragment thereof having serine protease activity as a single chain.

vii. The amount of experimentation required

There is nothing of record to suggest that production or use of any of the claimed polypeptides would require development of new procedures, techniques or excessive experimentation. Protein extraction, purification and synthesis methods have been used for decades. The specification provides a detailed working example for fermentation and isolated of an MTSP protease domain. As discussed above, MTSP family members are provided and described in the application and are well known in the art. The specification and the art describe conserved features that can be used to identify MTSP family members and the protease domain thereof. Such features include the catalytic triad, an N-terminal activation

cleavage site and conserved cysteines that participate in disulfide bonding. If needed, assays for evaluating activity of the polypeptides are taught in the specification and are known in the art. Such assays are routine in this art and do not require excessive experimentation.

The Examiner states that recombinant and mutagenesis techniques and enzyme isolation techniques are known and that it is routine to screen for multiple substitutions or multiple modifications as encompassed by the instant claims (see Final Office Action, Exhibit 1, page 11). As discussed, mutagenesis methods are not required to make and use the polypeptides as claimed. The instant claims are directed to isolated protease domains of MTSP family members; one of skill in the art can identify and isolate the protease domain of any MTSP family member, identify a free Cys and replace it with another amino acid as described in the application. Hence, the claimed polypeptides can be synthesized, isolated and characterized using routine testing, and, if necessary, one of skill in the art can test polypeptides for catalytic activity by routine experimentation using the assays provided in the specification or known to those of skill in art. Appellant notes that "a considerable amount of experimentation is permissible, if it is merely routine . . ." *In re Wands*, 858 F.3d 731, 737.

Conclusion

In light of the breadth of the claims, the extensive teachings and examples in the specification, the high level of skill of those in this art, the knowledge of those of skill in the art, and the fact identification and isolation of protease domains in MTSP family members and preparation of single chain forms thereof as well as variants thereof is predictable and reproducibly demonstrated, it would not require undue experimentation for one of skill in the art to make and use polypeptides with the features as claimed. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the isolated MTSP protease domains as claimed. Accordingly, Appellant respectfully submits that this rejection of claim 1 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

For the reasons above, each of the dependent claims meets the written description requirement and are enabled and, in addition, additional reasons for each dependent claim are described below.

Dependent Claim 11

Claim 11 depends from claim 1 and includes every limitation thereof. Claim 11 recites that the MTSP of the polypeptide of claim 1 is selected from among MTSP1, MTSP3, MTSP4

and MTSP6. The arguments set forth above with respect to claim 1 are incorporated herein. The specification describes MTSP1 and its protease domain, *e.g.*, at pages 54-58. The specification describes MTSP3 its protease domain, *e.g.*, at pages 58-60 and Example 1 (pages 160-167). The specification describes MTSP4 its protease domain, *e.g.*, at pages 60-63 and Example 2 (pages 167-171). The specification describes MTSP6 its protease domain, *e.g.*, at pages 63-64 and Example 3 (pages 171-176). The working examples demonstrate cloning of the protease domains, with replaced free Cys, for each of these.

In light of the breadth of the claims, the extensive teachings and examples in the specification, the high level of skill of those in this art, the knowledge of those of skill in the art, and the fact that it is predictable to identify protease domains in MTSP family members and prepare single chain forms thereof as well as variants thereof, it would not require undue experimentation for one of skill in the art to make and use polypeptides with the features as claimed. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the isolated MTSP protease domains of MTSP1, MTSP3, MTSP4 or MTSP6 of claim 11. Accordingly, Appellant respectfully submits that this rejection of claim 11 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 20

Claim 20 depends from claim 1 and includes every limitation thereof. The arguments set forth above with respect to claim 1 are incorporated herein. Claim 20 recites that the free Cys be replaced with a serine. The Examiner admits that recombinant and mutagenesis techniques are known in the art (see Final Office Action, Exhibit 2, page 11). The specification exemplifies the replacement of a free Cys in the protease domain with a serine residue. For example, see Example 1, which recites, on page 161, lines 4-9:

To eliminate the free cysteine (at position 310 in SEQ ID No. 4) that exists when the protease domain of the MTSP3 protein is expressed or the zymogen is activated, the free cysteine at position 310 (see SEQ ID No. 3), which is Cys122 if a chymotrypsin numbering scheme is used, was replaced with a serine.

Similarly the working Example provide MTSP4s, MTSP6 and MTSP1 with the free Cys replaced with serine. One of skill in the art readily can identify the protease domain of any MTSP family member, identify a free Cys and replace it with a serine residue. Such substitutions of amino acids are predictable and routine in the art.

In light of the breadth of claim 20, the extensive teachings and examples in the specification, the high level of skill of those in this art, the knowledge of those of skill in the

art, and the fact that it is predictable to replace a Cys with another amino acid residue, such as a serine residue, it would not require undue experimentation for one of skill in the art to make and use polypeptides with the features as claimed. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the isolated MTSP protease domains of claim 20. Accordingly, Appellant respectfully submits that this rejection of claim 20 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 34

Claim 34 depends from claim 1 and includes every limitation thereof. The arguments set forth above with respect to claim 1 are incorporated herein. Claim 34 recites the MTSP is selected from among corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4. For the reasons articulated above with respect to claim 1, Appellant respectfully submits that the specification is enabling for preparation and use of a substantially purified single-chain polypeptide consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, where the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain and a free Cys in the protease domain is replaced with another amino acid.

The specification specifically recites that the protease domains can be from any MTSP family member, including corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4. For example, see page 8, line 30 through page 10, line 2, which recites:

The protease domains provided herein include, but are not limited to, the single chain region having an N-terminus at the cleavage site for activation of the zymogen, through the C-terminus, or C-terminal truncated portions thereof that exhibit proteolytic activity as a single-chain polypeptide in in vitro proteolysis assays, of any MTSP family member, preferably from a mammal, including and most preferably human, that, for example, is expressed in tumor cells at different levels from non-tumor cells, and that is not expressed on an endothelial cell. These include, but are not limited to: MTSP1 (or matriptase), MTSP3, MTSP4 and MTSP6. Other MTSP protease domains of interest herein, particularly for use in in vitro drug screening proteolytic assays, include, but are not limited to: corin (accession nos. AF133845 and AB013874; see, Yan *et al.* (1999) J. Biol. Chem. 274:14926-14938; Tomia *et al.* (1998) J. Biochem. 124:784-789; Uan *et al.* (2000) Proc. Natl. Acad. Sci. U.S.A. 97:8525-8529; SEQ ID Nos. 61 and 62 for the human protein); enteropeptidase (also designated enterokinase; accession no. U09860 for the human protein; see, Kitamoto *et al.* (1995) Biochem. 27: 4562-4568; Yahagi *et al.* (1996) Biochem. Biophys. Res. Commun. 219:806-812; Kitamoto *et al.* (1994) Proc. Natl. Acad. Sci. U.S.A. 91:7588-7592; Matsushima *et al.* (1994) J. Biol. Chem. 269:19976-19982; see SEQ ID Nos. 63 and 64 for the human protein); human airway

trypsin-like protease (HAT; accession no. AB002134; see Yamaoka *et al.* J. Biol. Chem. 273:11894-11901; SEQ ID Nos. 65 and 66 for the human protein); hepsin (see, accession nos. M18930, AF030065, X70900; Yamaoka *et al.* (1988) J Biol Chem 27: 11895-11901; Vu *et al.* (1997) J. Biol. Chem. 272:31315-31320; and Farley *et al.* (1993) Biochem. Biophys. Acta 1173:350-352; SEQ ID Nos. 67 and 68 for the human protein); TMPRS2 (see, Accession Nos. U75329 and AF113596; Paoloni-Giacobino *et al.* (1997) Genomics 44:309-320; and Jacquinet *et al.* (2000) FEBS Lett. 468: 93-100; SEQ ID Nos. 69 and 70 for the human protein) TMPRSS4 (see, Accession No. NM 016425; Wallrapp *et al.* (2000) Cancer 60:2602-2606; SEQ ID Nos. 71 and 72 for the human protein); and TADG-12 (also designated MTSP6, see SEQ ID Nos. 11 and 12; see International PCT application No. WO 00/52044, which claims priority to U.S. application Ser. No. 09/261,416).

The application describes the protease domain of MTSP family members corin, MTSP1, enteropeptidase, HAT, TMPRSS4 and TMPRSS2. Each of the specified MTSP family members is known and characterized in the art. . In view of the instant application teaching that such protease domains have activity as single chain polypeptides, the skilled artisan can readily isolate the protease domain of any of corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4 as a single chain and replace the free Cys with another amino acid using routine techniques and if necessary test the isolated protease domain for the requisite activity.

Appellant respectfully submits that, in view of the arguments set forth above with respect to claim 1 and the teaching in the specification, which describes the MTSP family members corin, enteropeptidase, HAT, TMPRSS4 and TMPRSS2, the breadth of claim 34, the extensive teachings and examples in the specification, the high level of skill of those in this art, the knowledge of those of skill in the art, and the fact that it is predictable to isolate a protease domain and replace a Cys with another amino acid residue, it would not require undue experimentation for one of skill in the art to make and use polypeptides with the features of claim 34. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the isolated MTSP protease domains of claim 34. Accordingly, Appellant respectfully submits that this rejection of claim 34 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 35

Claim 35 is directed to a conjugate that includes a) a polypeptide of claim 1, and b) a targeting agent linked to the protein directly or via a linker, wherein the conjugate has serine protease activity. The arguments set forth above with respect to claim 1 are incorporated herein. The specification defines a "targeting agent" on page 38, lines 9-15, as:

any moiety, such as a protein or effective portion thereof, that provides specific binding of the conjugate to a cell surface receptor, which, preferably, internalizes the conjugate or MTSP portion thereof. A targeting agent may also be one that promotes or facilitates, for example, affinity isolation or purification of the conjugate; attachment of the conjugate to a surface; or detection of the conjugate or complexes containing the conjugate.

The specification teaches that the conjugates can be prepared by chemical conjugation, recombinant DNA technology or combinations thereof, and provides detailed descriptions of chemical conjugation, including acid cleavable, photo-cleavable and heat sensitive linker technology and other linkers, preparation of fusion proteins, peptide linkers, conjugation to targeting agents, and adsorption, absorption and/or covalent bonding to a solid support (see *e.g.*, pages 123-131). For example, the specification teaches that for the fusion proteins, the peptide or fragment thereof is linked to either the N-terminus or C-terminus of the MTSP protein domain (*e.g.*, see page 124, lines 25-26). The specification teaches that chemical conjugation also can be used to form conjugates, where the MTSP protein domain is linked via one or more selected linkers or directly to the targeting agent (*e.g.*, see page 126, lines 2-3). The specification describes various types of linkers and describes example of various linkers, including peptide linkers and chemical linkers, such as acid cleavable, photo-cleavable and heat cleavable linkers (*e.g.*, see pages 127-130). Methods of preparing protein conjugates are well known and routine in the art (*e.g.*, see Brinkley, "A Brief Survey of Methods for Preparing Protein Conjugates with Dyes, Haptens, and Cross-linking Reagents" in *Perspectives in Bioconjugate Chemistry* (Claude Meares, ed. 1993, Chapter 4, pages 59-70, Exhibit 6). Hence, routine techniques can be used to conjugate isolated protease domains to a targeting agent.

Appellant respectfully submits that, in view of the arguments set forth above with respect to claim 1 and the teaching in the specification, which describes conjugates of single-chain protease domains conjugated to a targeting agent, several different types of conjugation technologies for making the conjugates and exemplary conjugates, the breadth of claim 35, the high level of skill of those in this art, the knowledge of those of skill in the art, and the fact that it is routine and predictable to conjugate a polypeptide to a targeting agent, it would not require undue experimentation for one of skill in the art to make and use conjugates with the features of claim 35. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the conjugates of claim 35. Accordingly, Appellant respectfully submits that this rejection of claim 35 under

35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 36

Claim 36 depends from claim 35 and recites a conjugate that includes a targeting agent that permits i) affinity isolation or purification of the conjugate; ii) attachment of the conjugate to a surface; iii) detection of the conjugate; or iv) targeted delivery to a selected tissue or cell. The arguments set forth above with respect to claims 1 and 35 are incorporated herein.

The specification recites, I., at page 14, lines 19-26 and page 123, line 30 through page 124, line 7, that the targeting agent of the conjugate permits affinity isolation or purification of the conjugate; attachment of the conjugate to a surface; detection of the conjugate; or targeted delivery to a selected tissue or cell. The specification teaches exemplary targeting agents, including tissue specific or tumor specific monoclonal antibodies, a growth factor or fragment thereof, such as FGF, EGF, PDGF, VEGF, cytokines, including chemokines, and other such agents, a protein or peptide fragment that contains a protein binding sequence, a nucleic acid binding sequence, a lipid binding sequence, a polysaccharide binding sequence, or a metal binding sequence, or a linker for attachment to a solid support (see, I., page 124, lines 8-17 and pages 131-136). The specification also describes the construction of affinity binding pairs for isolation and/or purification of the conjugate (*e.g.*, see page 131, lines 5-37). Methods of preparing protein conjugates are well known and routine in the art (*e.g.*, see Brinkley, *supra*, Exhibit 6). Hence, routine, reproducible techniques well known to the skilled artisan can be used to conjugate isolated protease domains to a targeting agent.

Appellant respectfully submits that, in view of the arguments set forth above with respect to claims 1 and 35, and the teaching in the specification, which describes single-chain protease domains conjugated to a targeting agent and the use of such targeting agents for affinity isolation or purification of the conjugate or attachment of the conjugate to a surface or detection of the conjugate or targeted delivery to a selected tissue or cell, the breadth of claim 36, the high level of skill of those in this art, the knowledge of those of skill in the art, and the fact that it is routine and predictable to conjugate a polypeptide to a targeting agent, it would not require undue experimentation for one of skill in the art to make and use conjugates with the features of claim 36. Hence, a consideration of the factors enumerated

above leads to the conclusion that undue experimentation would not be required to make and use the isolated MTSP protease domains of claim 36. Accordingly, Appellant respectfully submits that this rejection of claim 36 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Dependent Claims 40 and 41

Claim 40 recites a solid support comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker. Claim 41 depends from claim 40 and recites that the polypeptides comprise an array. The arguments set forth above with respect to claim 1 are incorporated herein.

The specification describes solid supports and methods for immobilizing MTSP protein, such as a protease domain, to solid supports (*e.g.*, see pages 131-136). For example, the specification teaches exemplary solid supports, including supports having any required structure and geometry, such as beads, pellets, disks, capillaries, hollow fibers, needles, solid fibers, random shapes, thin films and membranes (*e.g.*, page 132, lines 26-29). The specification teaches that the solid support can be of any suitable material, such as inorganics, natural polymers, and synthetic polymers, including, cellulose, cellulose derivatives, acrylic resins, glass, silica gels, polystyrene, gelatin, polyvinyl pyrrolidone, co-polymers of vinyl and acrylamide, polystyrene cross-linked with divinylbenzene, polyacrylamides, latex gels, polystyrene, dextran, polyacrylamides, rubber, silicon, plastics, nitrocellulose, celluloses, natural sponges and highly porous glasses (*e.g.*, page 134, lines 1-30).

The specification teaches that a plurality of MTSP protease domains, including two or more protease domains, can be attached to a solid support (*e.g.*, page 132, lines 4-8). The instant specification defines an array as a collection of elements containing three or more members and that, as in the case for an addressable array, the members of the array can be immobilized to discrete identifiable loci on the surface of a solid phase (see, *e.g.*, page 35, lines 14-20).

The specification teaches that the polypeptide can be linked to the solid support directly or via a linker (*e.g.*, page 132, lines 1-2). The specification describes various linking technologies that can be used to link the polypeptide to the solid support (*e.g.*, page 135, lines 1-30). These include reacting the protein with a reactive moiety on the solid support and the specification describes exemplary reactive moieties, including amino silane linkages, hydroxyl linkages, carboxysilane linkages, N-[3-(triethoxy-silyl)propyl]phthelamic acid,

bis-(2-hydroxyethyl)aminopropyltriethoxysilane, derivatized polystyrenes (page 133, lines 7-26), absorption and adsorption or covalent binding to the support, either directly or via a linker, such as through disulfide linkages, thioether bonds, and covalent bonds between free reactive groups, such as amine and thiol groups, known to those of skill in art (page 135, lines 11-26). Linking a protein to a solid support is routine in the biotechnology arts (*e.g.*, see Means & Feeney, "Chemical Modifications of Proteins: History and Applications" in Perspectives in Bioconjugate Chemistry (Claude Meares, ed., 1993, Chapter 2, pages 10-20, Exhibit 23). The skilled artisan can select the appropriate conjugation chemistry based on the nature of the polypeptide and the solid support without undue experimentation and conjugate the protease domain to the solid support using routine techniques known in the art.

In light of the breadth of claims 40 and 41, the extensive teachings in the specification with respect to solid supports and conjugating polypeptides thereto, including conjugating a plurality of isolated protease domains to a solid support, the high level of skill of those in this art, and the knowledge of those of skill in the art, Appellant respectfully submits that it would not require undue experimentation for one of skill in the art to make and use the solid supports of claim 40 nor the arrays of claim 41. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the solid supports comprising two or more polypeptides of claim 40 linked thereto either directly or via a linker of claim 113 or the arrays of claim 41. Accordingly, Appellant respectfully submits that this rejection of claims 40 and 41 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 42

Claim 42 depends from claim 41 and recites that the array comprises polypeptides having different MTSP protease domains. The arguments set forth above with respect to claims 1, 40 and 41 are incorporated herein. The specification teaches that a plurality of MTSP protease domains can be attached to a solid support (*e.g.*, see page 132, lines 4-8). Linking a protein to a solid support is routine in the biotechnology arts (*e.g.*, see Means & Feeney, Chemical Modifications of Proteins: History and Applications in Perspectives in Bioconjugate Chemistry (Claude Meares, ed., 1993, Chapter 2, pages 10-20, Exhibit 23). Whether the protein to be conjugated to a solid support is a single species or multiple species of MTSP protease domain does not change the amount of experimentation required to form the claimed array. The skilled artisan readily can select the appropriate conjugation

chemistry based on the nature of the polypeptides and the solid support without undue experimentation and conjugate the polypeptide to the support using routine methods.

In light of the breadth of claim 42, the extensive teachings in the specification with respect to solid supports and conjugating polypeptides thereto, including conjugating a plurality of isolated protease domains to a solid support, the high level of skill of those in this art, and the knowledge of those of skill in the art, Appellant respectfully submits that it would not require undue experimentation for one of skill in the art to make and use the arrays of claim 42. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the arrays of claim 42. Accordingly, Appellant respectfully submits that this rejection of claim 42 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Dependent Claims 113 and 114

Claim 113 recites a solid support comprising two or more polypeptides of claim 12 linked thereto either directly or via a linker. Claim 114 depends from claim 113 and recites that the polypeptides comprise an array. Hence, each of claims 113 and 114 includes the polypeptide of claim 12 as an element. Claim 12 is not rejected under 35 U.S.C. §112, first paragraph. Accordingly, the Examiner admits that the specification is enabling for the subject matter of claim 12, which is directed to the substantially purified polypeptide of claim 1, wherein the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12.

The specification describes solid supports and methods for immobilizing MTSP protein to solid supports (*e.g.*, see pages 131-136). For example, the specification teaches exemplary solid supports, including supports having any required structure and geometry, such as beads, pellets, disks, capillaries, hollow fibers, needles, solid fibers, random shapes, thin films and membranes (*e.g.*, page 132, lines 26-29). The specification teaches that the solid support can be of any suitable material, such as inorganics, natural polymers, and synthetic polymers, including, cellulose, cellulose derivatives, acrylic resins, glass, silica gels, polystyrene, gelatin, polyvinyl pyrrolidone, co-polymers of vinyl and acrylamide, polystyrene cross-linked with divinylbenzene, polyacrylamides, latex gels, polystyrene, dextran,

polyacrylamides, rubber, silicon, plastics, nitrocellulose, celluloses, natural sponges and highly porous glasses (*e.g.*, page 134, lines 1-30).

The specification teaches that a plurality of MTSP protease domains, including two or more protease domains, can be attached to a solid support (*e.g.*, page 132, lines 4-8). The instant specification defines an array as a collection of elements containing three or more members and that, as in the case for an addressable array, the members of the array can be immobilized to discrete identifiable loci on the surface of a solid phase (see page 35, lines 14-20).

The specification teaches that the polypeptide can be linked to the solid support directly or via a linker (*e.g.*, page 132, lines 1-2). The specification describes various linking technologies that can be used to link the polypeptide to the solid support (*e.g.*, page 135, lines 1-30). These include reacting the protein with a reactive moiety on the solid support. The specification describes exemplary reactive moieties, including amino silane linkages, hydroxyl linkages, carboxysilane linkages, N-[3-(triethoxy-silyl)propyl]phthelamic acid and derivatized polystyrenes (page 133, lines 7-26). The specification also describes absorption and adsorption and covalent binding to the support, either directly or via a linker, such as via disulfide linkages or thioether bonds, and covalent bonds between free reactive groups, such as amine and thiol groups, known to those of skill in art (page 135, lines 11-26). Linking a protein to a solid support is routine in the biotechnology arts (*e.g.*, see Means & Feeney, *Chemical Modifications of Proteins: History and Applications in Perspectives in Bioconjugate Chemistry* (Claude Meares, ed., 1993, Chapter 2, pages 10-20, Exhibit 23). The skilled artisan readily can select the appropriate conjugation chemistry based on the nature of the polypeptides and the solid support without undue experimentation and conjugate the polypeptide to the support using routine methods.

In light of the breadth of claims 113 and 114, the extensive teachings in the specification with respect to solid supports and conjugating polypeptides thereto, including conjugating a plurality of isolated protease domains to a solid support, the high level of skill of those in this art, the knowledge of those of skill in the art, and the fact that the Examiner admits that the specification is enabling for the polypeptides of claim 12, Appellant respectfully submits that it would not require undue experimentation for one of skill in the art to conjugate the polypeptides of claim 12 to solid supports to make the solid supports of claim 113 and arrays of claim 114. Hence, a consideration of the factors enumerated above leads to the

conclusion that undue experimentation would not be required to make and use the solid supports comprising two or more polypeptides of claim 12 linked thereto either directly or via a linker of claim 113 or the arrays of claim 114. Accordingly, Appellant respectfully submits that this rejection of claims 113 and 114 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

Summary

In light of the breadth of the claims, the extensive teachings and examples in the specification, the high level of skill of those in this art, the knowledge of those of skill in the art, and the fact that it is predictable to identify protease domains in MTSP family members and prepare single chain forms thereof as well as variants thereof, it would not require undue experimentation for one of skill in the art to make and use polypeptides with the features as claimed, or conjugates, solid supports or arrays that include the polypeptides. Hence, a consideration of the factors enumerated above leads to the conclusion that undue experimentation would not be required to make and use the subject matter as claimed. Accordingly, Appellant respectfully submits that this rejection of claims 1, 11, 20, 34-36, 40-42, 113 and 114 under 35 U.S.C. §112, first paragraph, is erroneous in law and fact and, therefore, should be reversed.

3. REJECTION OF CLAIMS 1, 11-13, 20, 34-36, 40-42, 113 AND 114 UNDER 35 U.S.C.

§102(b) - Takeuchi

Claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 are rejected under 35 U.S.C. §102(b) as being anticipated by Takeuchi, because the reference allegedly discloses "a polypeptide comprising a fragment consisting of a serine protease domain that is 100% identical to amino acids 615-855 of SEQ ID NO:2 of the instant invention" and discloses "a catalytically active polypeptide comprising the serine protease domain linked to a His-tag." The Examiner states that Takeuchi discloses that Cys at position 731 forms a disulfide bond with Cys 604 present in the pro domain (see Final Office Action, Exhibit 2, page 17). The Examiner alleges that the claim limitation "a free Cys in the protease domain is replaced with another amino acid" and "a free Cys in the protease domain is replaced with a serine" is a product-by-process type limitation. The Examiner alleges that

[t]he end result of the products of the claims is a serine protease domain or a serine protease domain having a serine residue. Whether the product of the claimed protein is obtained by replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not. Therefore, there is no there a structure implied by

said limitations. Since the polypeptide of Takeuchi *et al.* consists of a protease domain of a MTSP and the MTSP protease domain has serine protease activity, the claims are anticipated by the prior art. Also, since the serine protease domain of Takeuchi *et al.* has a serine residue, claim 20 is also anticipated.

The rejection respectfully is traversed.

A. LEGAL STANDARDS - ANTICIPATION UNDER 35 U.S.C. § 102

Anticipation is a factual determination that "...requires the presence in a single prior art disclosure of each and every element of a claimed invention." *Lewmar Marine, Inc. v. Barient, Inc.*, 3 U.S.P.Q.2d 1766 (Fed. Cir. 1987). Moreover, "[a] claim is anticipated only if each and every element as set forth in the claim is found, either expressly or inherently described, in a single prior art reference." *Verdegaal Bros. v. Union Oil of California*, 2 U.S.P.Q.2d 1051, 1053 (Fed. Cir. 1987) (emphasis added).

Federal Circuit decisions have repeatedly emphasized the notion that anticipation cannot be found where less than all elements of a claimed invention are set forth in a reference. See, e.g. *Transclean Corp. v. Bridgewood Services, Inc.*, 290 F.3d 1364 (Fed. Cir. 2002). In this regard, a reference disclosing "substantially the same thing" is not enough to anticipate. *Jamesbury Corp. v. Litton Indust. Prod., Inc.*, 756 F.2d 1556, 1560 (Fed. Cir. 1985). A reference must clearly disclose each and every limitation of the claimed invention before anticipation may be found.

Further, anticipation cannot be shown by combining more than one reference to show the elements of the claimed invention. In *re Saunders*, 444 F.2d 599 (C.C.P.A. 1971). All elements of a claimed invention must be disclosed in one, solitary reference. As such, it is clear that a reference cannot be utilized to render a claimed invention anticipated without identical disclosure.

B. THE REJECTION OF CLAIMS 1, 11-13, 20, 34-36, 40-42, 113 AND 114 UNDER 35 U.S.C. §102(b) SHOULD BE REVERSED BECAUSE TAKEUCHI DOES NOT ANTICIPATE THE CLAIMED SUBJECT MATTER

1. Disclosure of Takeuchi

Takeuchi discloses a polypeptide that contains 855 amino acids and is designated MT-SP1. This protein has sequence identity with the full-length MTSP1 set forth as SEQ ID NO:2 of the instant application. Takeuchi discloses an expression vector that includes nucleic acid encoding the protease domain plus the pro-domain (see page 11055, left col., third full paragraph). Takeuchi discloses that its expression vector includes the mature protease domain and a small portion of the pro-domain and was designed to over-express the sequence encoding

a polypeptide containing amino acids 596-855 with a His-tag fusion to produce as a construct Met-Arg-Gly-Ser-His₆-aa596-855 (page 11055, column 2, third full paragraph). Takeuchi discloses that amino acids Cys 604 and Cys 731 are disulfide bonded (see for example, at page 11060, col. 1). Takeuchi discloses that its protease domain is disulfide bonded to the pro-domain region (see page 11055, column 2, third full paragraph and page 11058, col. 1 and page 11060, col. 1, first paragraph) and that the pro-domain region remains bonded to the protease domain after activation (page 11058, lines 8-9).

Takeuchi discloses that its "purified protease domain" includes the His-tag sequence and the pro-domain region bonded thereto, stating that a monoclonal antibody directed against the N-terminal Arg-Gly-Ser-His₄ epitope is immunoreactive with its purified protein (see page 11058). It is **not** an isolated single chain protease domain. It is a two chain structure and it includes amino acids in addition to the protease domain. Figure 3 cited by the Examiner as showing an isolated protease domain is a diagrammatic representation of the MTSP1 protease domains; it by no means is an isolated protease domain. Furthermore, the figure depicts the disulfide bonds and does not show a free Cys in the protease domain, nor a fragment consisting of the protease domain. Page 11057, referenced by the Examiner as describing isolation of protease domain, does not do so. The polypeptide is expressed as a His-tagged polypeptide that **forms a two-chain structure** by virtue of the Cys-Cys disulfide bonds depicted in Figure 3. Furthermore, the paper discusses the activated His-tag extended polypeptide and describes its activity (see, e.g., Figure 6 and page 11057, col. 2). Takeuchi states that:

the MT-SP1 protease domain was expressed in *E. coli* as a His-tagged fusion and was purified from inclusion bodies under denaturing conditions by using metal-chelate affinity chromatography. . . . This denatured protein refolded when the urea was dialyzed from the protein. . . . N-terminal sequencing of the purified activated [i.e. the two-chain folded form] yielded the expected VVGGT activation sequence.

Thus, Takeuchi expresses a His-tagged form of the protein, which includes a protease domain **and** a pro-domain region, that forms a two chain structure when activation- cleaved. The sequenced molecule includes the His-tagged protease domain. Takeuchi does not disclose or contemplate an isolated polypeptide consisting of only the protease domain and does not mention replacement of any Cys with Ser (the Cys in its two-chain form is **not** free).

Further, it is apparent from the disclosure that Takeuchi believes that a two-chain structure is a requisite for activity. Takeuchi discusses the need for activation cleavage and depicts the disulfide bond; there is no disclosure of a polypeptide in which there is a free Cys.

Hence, there is no disclosure for replacing any free Cys with another amino acid, such as a serine. There is no mention of replacement of any amino acids in its polypeptide.

Hence Takeuchi does not disclose isolation of a polypeptide consisting only of the protease domain of any MTSP, including an MTSP1. Its polypeptide includes a His-tag sequence; the active form of the enzyme includes a disulfide bond between the protease domain and a pro-domain region. In addition, the only isolation of a polypeptide including the protease domain (which includes the His-tag), was for sequencing purposes.

2. Analysis

Independent Claim 1

In maintaining the rejection, the Examiner states on page 18 of the Final Office Action (Exhibit 2) that:

[t]he limitation "a free Cys in the protease domain is replaced with another amino acid" and "a free Cys in the protease domain is replaced with a serine" is a product-by-process type limitation. The end result of the products of the claims is a serine protease domain or a serine protease domain having a serine residue. Whether the product of the claimed protein is obtained by replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not. Therefore, there is no [] structure implied by said limitations. Since the polypeptide of Takeuchi *et al.* consists of a protease domain of a MTSP and the MTSP protease domain has serine protease activity, the claims are anticipated by the prior art. Also, since the serine protease domain of Takeuchi *et al.* has a serine residue, claim 20 is also anticipated.

Appellant respectfully disagrees. Claim 1 recites that the isolated substantially purified polypeptide consists only of a protease domain or a smaller catalytically active portion of the protease as a single chain, and that a free Cys residue of the serine protease domain is replaced with another amino acid. This is not a "product-by-process type" limitation as alleged by the Examiner, but a limitation on the molecular structure of the single chain polypeptide.

A product-by-process claim is a product claim that defines the claimed product in terms of the process by which it is made. *In re Luck*, 476 F.2d 650, 177 USPQ 523 (CCPA 1973); *In re Pilkington*, 411 F.2d 1345, 162 USPQ 145 (CCPA 1969); *In re Steppan*, 394 F.2d 1013, 156 USPQ 143 (CCPA 1967). Appellant respectfully submits that the instant claims do not define the product in terms of the process by which it is made. The specification teaches that a single-chain form of a serine protease domain has a free Cys residue. For example, page 58, lines 12-20 recites:

Muteins of the MTSP1 proteins are provided. In the activated double chain molecule, residue 731 forms a disulfide bond with the Cys at residue 604. In the single chain form, the residue at 731 in the protease domain is free. Muteins in which Cys residues, particularly the free Cys residue (amino acid 731 in SEQ ID No. 2) in the single chain protease domain [is replaced] are provided. Other muteins in which conservative amino

acids replacements are effected and that retain proteolytic activity as a single chain are also provided. Such changes may be systematically introduced and tested for activity in in vitro assays, such as those provided herein.

The Cys residue in the protease domain in the MTSP protein forms a disulfide bond with a Cys residue in pro-domain region, and autoactivation results in a polypeptide with a two-chain structure by virtue of the Cys–Cys disulfide bonds. Isolating the serine protease domain so that it is free from the pro-domain region results in unpaired Cys residues, because the single-chain isolated protease domain is not bonded to a Cys in another region of the protein, such as the pro-domain region. Hence, the isolated polypeptide consisting only of the protease domain will have a free Cys residue (a Cys residue that “does not form disulfide linkages with any other Cys residue in the protein,” see page 10, lines 5-6 of the instant specification). Thus, the isolation of the protease domain results in a free Cys residue. Isolation of the protease domain does not result in a free Cys residue that is replaced with another amino acid. Further, the single chain form of the single chain protease domain can be made by recombinant expression in a vector, thus eliminating the need to “isolate” it from the expressed zymogen form of the enzyme. The isolated single chain form of the serine protease domain is not produced by replacing a free Cys residue with another amino acid. Hence, the claimed polypeptide is not defined in terms of the process by which it was made. Accordingly, the instant claims are not “product-by-process” claims. The polypeptides of Takeuchi *et al.* are two-chain polypeptides and do not contain a free Cys; hence they cannot contain a replaced free Cys.

The limitation a free Cys residue of the serine protease domain is replaced with another amino acid is a structural limitation on the molecular architecture of the polypeptide. Cys residues readily form disulfide bonds due to the presence of the sulfhydryl group (*e.g.*, see Zubay, *Biochemistry* ((1983), pages 12-13, Exhibit 45). Other amino acid residues do not have this functionality. For example, serine residues have a hydroxyl group instead of a sulfhydryl group and thus do not form disulfide bonds. Hence, replacing a free Cys residue in the protease domain of the polypeptide with another amino acid, such as a serine residue, as is claimed in claim 20, results in a protease domain that cannot form a disulfide bond with another region in the polypeptide. Hence, the recited limitation is a structural limitation. If the claims recited “wherein a sulfhydryl group is replaced with another functionality” instead of “wherein a free Cys residue of the serine protease domain is replaced with another amino acid” there would be no question that the recitation is a structural limitation on the claimed compound. Because the recitation limits the structure of the polypeptide, the recited limitation

a free Cys residue of the serine protease domain is replaced with another amino acid should be afforded patentable weight. "All words in a claim must be considered in judging the patentability of that claim against the prior art." In re Wilson, 424 F.2d 1382, 1385, 165 USPQ 494, 496 (CCPA 1970).

Appellant respectfully submits that Takeuchi does not disclose every element of the claimed subject matter.

(1) Free Cys residue

Takeuchi does not disclose a serine protease domain of an MTSP polypeptide that has a free Cys residue. Figure 3 of Takeuchi, for example, is a diagrammatic representation of the full-length MTSP1 depicting the activated disulfide-bonded form of the enzyme, in which the Cys residue of the protease domain is part of a disulfide bond with a Cys residue in the pro-domain. Figure 4 of Takeuchi, which shows multiple sequence alignments of MTSP1 structural motifs, identifies Cys residues that participate in disulfide bonds. All of the Cys residues in Figure 4 are shown as being disulfide bonded – there are no free Cys residues. Takeuchi discloses that its protease domain is disulfide bonded to the pro-domain region and remains bonded to the protease domain after activation and thus Takeuchi does not disclose a protease domain having a free Cys residue.

(2) Replacing a free Cys residue with another amino acid

There is no disclosure in Takeuchi with respect to replacement of any amino acid in its polypeptide. Takeuchi does not disclose replacing any amino acid in the serine protease domain with another amino acid. As discussed above, Takeuchi does not disclose a serine protease domain of an MTSP polypeptide that has a free Cys residue. Hence, Takeuchi does not disclose replacing a free Cys residue of the serine protease domain of an MTSP polypeptide with another amino acid.

The Examiner's argument that "the serine protease domain of Takeuchi has a serine residue" and thus "claim 20 is also anticipated" is incorrect. Claim 20 does not recite a serine protease domain that has a serine residue. The claims recite that a free Cys residue of the serine protease domain of an MTSP polypeptide is replaced with another amino acid. There is no disclosure in Takeuchi of a protease domain of an MTSP polypeptide having a free Cys residue of the serine protease domain replaced with another amino acid. It is irrelevant whether other amino acid residues in the protease domain are serine residues.

3) An isolated, substantially purified protease domain of an MTSP polypeptide

Takeuchi discloses that its protease domain is disulfide bonded to the pro-domain region and remains bonded to the protease domain after activation. Takeuchi discloses that its “purified protease domain” includes the His-tag sequence, and states that a monoclonal antibody directed against the N-terminal Arg-Gly-Ser-His₄ epitope is immunoreactive with its purified protein. Thus, the “purified protease domain” disclosed by Takeuchi includes additional amino acid residues in addition to the protease domain of the MTSP1. Neither page 11057 nor Figure 3 of Takeuchi discloses a single chain polypeptide that consists only of the protease domain. As discussed above, the protease domain as expressed and isolated by Takeuchi includes additional amino acids. Takeuchi states that:

N-terminal sequencing of the purified activated [i.e. the two-chain folded form] yielded the expected VVGGT activation sequence.

The purified activated polypeptide according to Takeuchi is a two chain polypeptide, and also, as expressed, includes the His-tag for purification. Figure 3, as noted, is a diagrammatic representation of the full-length MTSP1 depicting the activated disulfide-bonded form of the enzyme (in which the Cys that is replaced in the instant claims, is part of the disulfide bond). Hence, Takeuchi does not disclose a polypeptide consisting only of a protease domain or a smaller catalytically active portion of the protease domain. Thus, Takeuchi does not disclose an isolated, substantially purified protease domain of an MTSP polypeptide having a free Cys residue replaced with another amino acid. Hence, the disclosure of Takeuchi does not disclose every element of claim 1. Therefore, Takeuchi does not anticipate claim 1 nor any claim dependent thereon. Accordingly, Appellant respectfully submits that the rejection of claim 1 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

For the reasons above, Takeuchi does not anticipate any of the dependent claims and, in addition, additional reasons why Takeuchi does not anticipate each dependent claim are described below.

Dependent Claim 11

Claim 11 depends from claim 1 and recites that the MTSP is selected from among MTSP1, MTSP3, MTSP4 and MTSP6. Claim 11 includes every limitation of claim 1, from which it depends. For the reasons discussed above with respect to claim 1, Takeuchi does not disclose every element of claim 11 and therefore does not anticipate claim 11. Accordingly,

Appellant respectfully submits that the rejection of claim 11 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 12

Claim 12 depends from claim 1 and recites that the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2 (MTSP1 protease domain), amino acids 205-437 of SEQ ID NO. 4 (MTSP3), the amino acid residues set forth as SEQ ID No. 6 (MTSP4) or as amino acids 217-443 in SEQ ID No. 12 (MTSP6), where the free Cys is replaced with Ser. Claim 12 includes every limitation of claim 1, from which it depends. For the reasons discussed above with respect to claim 1, Takeuchi does not disclose every element of claim 12 and therefore does not anticipate claim 12. Accordingly, Appellant respectfully submits that the rejection of claim 12 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 13

Claim 13 depends from claim 1 and recites that the substantially purified polypeptide has at least about 95% sequence identity with a protease domain consisting of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acids set forth as SEQ ID No. 6, and amino acids 217-443 in SEQ ID No. 12. Claim 13 includes every limitation of claim 1, from which it depends. For the reasons discussed above with respect to claim 1, Takeuchi does not disclose every element of claim 13 and therefore does not anticipate claim 13. Accordingly, Appellant respectfully submits that the rejection of claim 13 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 20

Claim 20 depends from claim 1 and recites that a free Cys in the protease domain is replaced with a serine. Claim 20 includes every limitation of claim 1, from which it depends. As discussed above, Takeuchi does not disclose a serine protease domain of an MTSP polypeptide that has a free Cys residue. There is no disclosure in Takeuchi with respect to replacement of any amino acids in its polypeptide. Takeuchi does not disclose replacing any amino acid in the serine protease domain with another amino acid. Takeuchi does not disclose replacing a free Cys residue of the serine protease domain of an MTSP polypeptide with a serine. Thus, for these reasons and the reasons discussed above with respect to claim 1, Takeuchi does not disclose every element of claim 20 and therefore does not anticipate claim

20. Accordingly, Appellant respectfully submits that the rejection of claim 20 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 34

Claim 34 depends from claim 1 and recites that the MTSP is selected from among corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4. Claim 34 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, Takeuchi does not disclose every element of claim 34 and therefore does not anticipate claim 34. Accordingly, Appellant respectfully submits that the rejection of claim 34 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 40

Claim 40 recites a solid support comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker. Takeuchi does not disclose an isolated single-chained polypeptide consisting only of an MTSP protease domain in which a free Cys has been replaced with another amino acid nor conjugating two or more such isolated protease domains to a solid support. Hence, there is no disclosure in Takeuchi of a solid support that includes two or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Thus, for these reasons and the reasons discussed above with respect to claim 1, Takeuchi does not disclose every element of claim 40 and therefore does not anticipate claim 40. Accordingly, Appellant respectfully submits that the rejection of claim 40 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 41

Claim 41 recites a solid support comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker where the polypeptides comprise an array. The specification defines an array as a collection of elements containing three or more members. As discussed above, Takeuchi does not disclose isolating the protease domain and preparing it as a single chain and modifying the single-chain polypeptide that has a free Cys residue by replacing the free Cys residue with another amino acid. Takeuchi does not disclose a solid support that includes three or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Thus, for these reasons and the reasons discussed above with respect to claim 1, Takeuchi does not

disclose every element of claim 41 and therefore does not anticipate claim 41. Accordingly, Appellant respectfully submits that the rejection of claim 41 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 42

Claim 42 depends from claim 41 and recites that the array comprises polypeptides having different MTSP protease domains. As discussed above, Takeuchi does not disclose isolating the protease domain and preparing it as a single chain nor replacing any amino acid in the MTSP polypeptide with another amino acid. Takeuchi does not disclose modifying a single-chain polypeptide that has a free Cys residue by replacing the free Cys residue with another amino acid. Takeuchi does not disclose a solid support that includes three or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Takeuchi does not disclose a solid support that includes three or more isolated protease domains in which a free Cys was replaced with another amino acid, where the protease domains are from different MTSPs. Thus, for these reasons and the reasons discussed above with respect to claim 1, Takeuchi does not disclose every element of claim 42 and therefore does not anticipate claim 42. Accordingly, Appellant respectfully submits that the rejection of claim 42 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 113

Claim 113 recites a solid support comprising two or more polypeptides of claim 12 linked thereto either directly or via a linker. Claim 12 depends from claim 1 and specifies that the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12. Claim 12 includes every limitation of claim 1, from which it depends.

Takeuchi does not disclose isolating the protease domain and preparing it as a single chain. Takeuchi does not disclose replacing any amino acid in the MTSP polypeptide with another amino acid, and does not disclose modifying a single-chain polypeptide that has a free Cys residue by replacing the free Cys residue with another amino acid. There is no disclosure in Takeuchi of a solid support that includes two or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Thus, for these reasons and the reasons discussed above with respect to claim 1

and claim 12, Takeuchi does not disclose every element of claim 113 and therefore does not anticipate claim 113. Accordingly, Appellant respectfully submits that the rejection of claim 113 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 114

Claim 114 depends from claim 113 and recites that the polypeptides comprise an array. As discussed above, Takeuchi does not disclose isolating the protease domain and preparing it as a single chain. Takeuchi does not disclose replacing any amino acid in the MTSP polypeptide with another amino acid, and does not disclose modifying a single-chain polypeptide that has a free Cys residue by replacing the free Cys residue with another amino acid. There is no disclosure in Takeuchi of a solid support that includes three or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Thus, for these reasons and the reasons discussed above with respect to claim 1 and claim 113, Takeuchi does not disclose every element of claim 114 and therefore does not anticipate claim 114. Accordingly, Appellant respectfully submits that the rejection of claim 114 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Summary

Appellant respectfully submits that, in light of the above, the Examiner has failed to establish claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 as anticipated by Takeuchi under 35 U.S.C. §102(b). Accordingly, Appellant respectfully submits that the rejection of claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

. THE REJECTION OF CLAIMS 1, 11-13 AND 34 UNDER 35 U.S.C. §102(e)/103(a)

In the Final Office Action (Exhibit 1), on page 19, claims 1, 11-13 and 34 are rejected as obvious under 35 U.S.C. §103(a) over O'Brien and there is no mention of a rejection under 35 U.S.C. §102(e), although the rejection is set forth under the heading "Claim Rejections - 35 USC §102/103." In the paragraph bridging pages 20 and 21 of the Final Office Action, however, the Examiner states that the claims are anticipated by O'Brien. Accordingly, Appellant separately traverses the rejection of claims 1, 11-13 and 34 under 35 U.S.C. §102(e) as anticipated by O'Brien and the rejection of claims 1, 11-13 and 34 as obvious under 35 U.S.C. §103(a) over O'Brien.

THE 102(e) REJECTION

The Examiner alleges that the limitation “a free Cys residue of the serine protease domain is replaced with another amino acid” is a “product-by-process type” limitation, and that “whether the product is obtained by replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not” and concludes that “there is no structure implied by said limitations. The Final Office Action concludes that the disclosed molecules in O’Brien anticipate the claimed subject matter.

A. LEGAL STANDARDS - ANTICIPATION UNDER 35 U.S.C. § 102(b)

The law with respect to anticipation under 35 U.S.C. § 102(a) is discussed above.

B. THE REJECTION OF CLAIMS 1, 11-13 AND 34 UNDER 35 U.S.C. §102(b) SHOULD BE REVERSED BECAUSE O’BRIEN DOES NOT ANTICIPATE THE CLAIMED SUBJECT MATTER

1. The disclosure of O’Brien

O’Brien discloses a protein identified therein as TADG-15, which is an MTSP1 variant, with a sequence of amino acids as set forth as SEQ ID NO:2. The reference also discloses a comparison of the amino acid sequence of the protease domain of TADG-15 (SEQ ID NO:14) with other serine protease catalytic domains (see Figure 2). O’Brien discloses that TADG-15 is a highly over-expressed gene in tumors and suggests that TADG-15 is novel in its component structure of domains because it has a protease catalytic domain that could be released *in vivo* and used as a diagnostic *in vivo* and that potentially could be a target for therapeutic intervention (col. 15, lines 31-38):

TADG-15 is a highly overexpressed gene in tumors. It is expressed in a limited number of normal tissues, primarily tissues that are involved in either uptake or secretion of molecules e.g. colon and pancreas. TADG-15 is further novel in its component structure of domains in that it has a protease catalytic domain which could be released and used as a diagnostic and which has the potential for a target for therapeutic intervention.

Thus, O’Brien states that the TADG-15 protease domain possibly could be released *in vivo* and serve as a therapeutic target, **not** as a therapeutic. O’Brien does not disclose, teach or suggest or mention or even hint at isolating the protease domain nor provide any disclosure that isolation of a protease domain would result in a free Cys that should be replaced.

O’Brien does not disclose isolation of the protease domain as a single-chain polypeptide that consists only of the protease domain as a single chain. O’Brien does not disclose a protease domain of an MTSP polypeptide that has a free Cys residue, or replacing

a free Cys residue of a serine protease domain of an MTSP polypeptide with another amino acid.

2. ANALYSIS

Independent Claim 1

Claim 1 recites that the isolated substantially purified polypeptide consists of a protease domain or a smaller catalytically active portion of the protease as a single chain, and that a free Cys residue of the serine protease domain is replaced with another amino acid. O'Brien does not disclose an isolated polypeptide that consists only of a protease domain or a smaller catalytically active portion of the protease as a single chain. O'Brien does not disclose an isolated single-chain protease domain of an MTSP polypeptide having a free Cys residue, or replacing a free Cys residue of an isolated single-chain serine protease domain of an MTSP polypeptide with another amino acid. In the previous Office Action, mailed April 21, 2006 (Exhibit 46, at page 20, lines 6-7), the Examiner states that O'Brien does not disclose a protease domain that has been purified. Hence, O'Brien does not disclose every element of claim 1.

In addition, as discussed above, O'Brien does not disclose an isolated protease domain of an MTSP. Stating that such protease domain could be released *in vivo* and used as a diagnostic target does not constitute a disclosure of an isolated single chain protease domain, and certainly does not constitute disclosure of an isolated protease domain in which a free Cys is replaced.

In maintaining the rejection, the Examiner states on page 20 of the Final Office Action (Exhibit 1) that

[t]he limitation "a free Cys in the protease domain is replaced with another amino acid" is a product-by-process type limitation. The end result of the products of the claims is a serine protease domain. Whether the product of the claimed protein is obtained by replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not. Therefore, there is no there a structure implied by said limitations. Since the polypeptide of O'Brien *et al.* consists of a protease domain of a MTSP and the MTSP protease domain has serine protease activity, the claims are anticipated by the prior art.

Appellant respectfully submits that a free Cys residue of the serine protease domain is replaced with another amino acid is not a "product-by-process type" limitation as alleged by the Examiner, but a limitation on the molecular structure of the single chain polypeptide. A product-by-process claim is a product claim that defines the claimed product in terms of the process by which it is made. Appellant respectfully submits that the instant claims do not

define the product in terms of the process by which it is made. As taught in the specification (e.g., see page 58, lines 12-20, which is reproduced above in the traverse of the rejection over Takeuchi), the Cys residue in the protease domain in the MTSP protein forms a disulfide bond with a Cys residue in pro-domain region, and autoactivation results in a polypeptide with a two-chain structure by virtue of the Cys–Cys disulfide bonds. Isolating the serine protease domain so that it is free from the pro-domain region results in unpaired Cys residues, because the Cys residue in the protease domain of the single-chain isolated protease domain is not bonded to a Cys in another region of the protein, such as the pro-domain region. Thus, the isolated polypeptide consisting only of the protease domain will have a free Cys residue (a Cys residue that “does not form disulfide linkages with any other Cys residue in the protein,” see page 10, lines 5-6 of the instant specification). Thus, the isolation of the protease domain results in a free Cys residue. Isolation of the protease domain does not result in a free Cys residue being replaced with another amino acid. Further, the single chain form of the single chain protease domain can be made by recombinant expression in a vector, thus eliminating the need to “isolate” it from the expressed zymogen form of the enzyme. The isolated single chain form of the serine protease domain is not produced by replacing a free Cys residue. Hence, the claimed polypeptide is not defined in terms of the process by which it was made. Accordingly, the instant claims are not “product-by-process” claims.

The limitation a free Cys residue of the serine protease domain is replaced with another amino acid is a structural limitation on the molecular architecture of the polypeptide. Cys residues readily form disulfide bonds due to the presence of the sulfhydryl group (e.g., see Zubay, Biochemistry ((1983), pages 12-13, Exhibit 45). Other amino acid residues do not have this functionality. For example, serine residues have a hydroxyl group instead of a sulfhydryl group and thus do not form disulfide bonds. Hence, replacing a free Cys residue in the protease domain of the polypeptide with another amino acid, such as a Ser residue, as is claimed in claim 20, results in a protease domain that cannot form a disulfide bond with another region in the polypeptide. Hence, the recited limitation is a structural limitation. Because the recitation limits the structure of the polypeptide, the recitation should be afforded patentable weight. “All words in a claim must be considered in judging the patentability of that claim against the prior art.” In re Wilson, 424 F.2d 1382, 1385, 165 USPQ 494, 496 (CCPA 1970).

Hence, O'Brien does not disclose every element of claim 1. Therefore O'Brien does not anticipate claim 1 nor any claim dependent thereon. Accordingly, Appellant respectfully submits that the rejection of claim 1 as anticipated by O'Brien is erroneous in law and fact and, therefore, should be reversed.

For the reasons above, O'Brien does not anticipate any of the dependent claims and, further, additional reasons why O'Brien does not anticipate each dependent claim are described below.

Dependent Claim 11

Claim 11 depends from claim 1 and specifies that the MTSP is selected from among MTSP1, MTSP3, MTSP4 and MTSP6. Claim 11 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien does not disclose every element of claim 11 and therefore does not anticipate claim 11. Accordingly, Appellant respectfully submits that the rejection of claim 11 as anticipated by O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 12

Claim 12 depends from claim 1 and specifies that the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12. Claim 12 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien does not disclose every element of claim 12 and therefore does not anticipate claim 12. Accordingly, Appellant respectfully submits that the rejection of claim 12 as anticipated by O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 13

Claim 13 depends from claim 1 and specifies that the substantially purified polypeptide has at least about 95% sequence identity with a protease domain consisting of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acids set forth as SEQ ID No. 6, and amino acids 217-443 in SEQ ID No. 12. Claim 13 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien does not disclose every element of claim 13 and therefore does not anticipate claim 13. Accordingly, Appellant

respectfully submits that the rejection of claim 1 as anticipated by O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 34

Claim 34 depends from claim 1 and specifies that the MTSP is selected from among corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4. Claim 34 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien does not disclose every element of claim 34 and therefore does not anticipate claim 34. Accordingly, Appellant respectfully submits that the rejection of claim 1 as anticipated by O'Brien is erroneous in law and fact and, therefore, should be reversed.

Summary

Appellant respectfully submits that, in light of the above, the Examiner has failed to establish claims 1, 11-13 and 34 as anticipated under 35 U.S.C. §102(b) by O'Brien. Accordingly, Appellant respectfully submits that the rejection of claims 1, 11-13 and 34 as anticipated by O'Brien is erroneous in law and fact and, therefore, should be reversed.

5. THE REJECTION OF CLAIMS 1, 11-13 AND 34 AND CLAIMS 35, 36, 40-42, 113 AND 114 UNDER 35 U.S.C. §103(a) – O'Brien

Claims 1, 11-13 and 34, as well as claims 35, 36, 40-42, 113 and 114, are rejected as unpatentable over O'Brien under 35 U.S.C. §103(a) because O'Brien allegedly teaches a method of expressing polypeptides in host cells and that it teaches that the protease domain could be released from the polypeptide and used as a diagnostic that has the potential for therapeutic intervention. Thus, the Final Office Action concludes that it would have been obvious to one of skill in the art to express the protease domain disclosed as SEQ ID NO:14 by O'Brien and purify the polypeptide. It is alleged that the motivation to make such polypeptides is the disclosed use as a diagnostic for therapeutic intervention. Further, it is alleged that one of ordinary skill in the art would have had a reasonable expectation of success since the expression of heterologous polypeptides was routine in the art and O'Brien teaches how to express heterologous polypeptides. The Examiner also alleges that the limitation "a free Cys residue of the serine protease domain is replaced with another amino acid" is a "product-by-process type" limitation, and that "whether the product is obtained by replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not" and concludes that "there is no structure implied by said limitations.

The rejection respectfully is traversed. As discussed above, O'Brien *et al.* speculates that the protease domain of TAG-15 could be released *in vivo* and, if it turns out that it is released *in vivo*, the protease domain could serve as therapeutic target. This is not a teaching or suggestion or even hint for producing the protease domain *in vitro* and using it as a therapeutic (not a target) or as a diagnostic reagent)not as a target. There is nothing taught or suggested in O'Brien *et al.* would have led one of ordinary skill in the art to isolate the protease domain (or a catalytically active fragment there) and replace what ends up as a free Cys with another amino acid.

A. LEGAL STANDARDS - OBVIOUSNESS UNDER 35 U.S.C. § 103(a)

For prima facie obviousness of claimed subject matter to be established under 35 U.S.C. §103, all the claim limitations must be taught or suggested by the prior art. In re Royka, 490 F.2d 981, 180 USPQ 580 (CCPA 1974). This principle of U.S. law regarding obviousness was not altered by the recent Supreme Court holding in KSR International Co. v. Teleflex Inc., 127 S.Ct. 1727, 82 USPQ2d 1385 (2007). In KSR, the Supreme Court stated that "Section 103 forbids issuance of a patent when 'the differences between the subject matter sought to be patented and the prior art are such the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains.'" KSR Int'l Co. v. Teleflex Inc., 127 S.Ct. 1727, 1734, 82 USPQ2d 1385, 1391 (2007).

The mere fact that prior art may be modified to produce the claimed product does not make the modification obvious unless the prior art suggests the desirability of the modification. In re Fritch, 23 U.S.P.Q.2d 1780 (Fed. Cir. 1992); see, also, In re Papesch, 315 F.2d 381, 137 U.S.P.Q. 43 (CCPA 1963). Further, that which is within the capabilities of one skilled in the art is not synonymous with that which is obvious. Ex parte Gerlach, 212 USPQ 471 (Bd. APP. 1980).

Furthermore, the Supreme Court in KSR took the opportunity to reiterate a second long-standing principle of U.S. law: that a holding of obviousness requires the fact finder (here, the Examiner), to make explicit the analysis supporting a rejection under 35 U.S.C. 103, stating that "rejections on obviousness cannot be sustained by mere conclusory statements; instead, there must be some articulated reasoning with some rational underpinning to support the legal conclusion of obviousness. Id. at 1740-41, 82 USPQ2d at 1396 (citing In re Kahn, 441 F.3d 977, 988, 78 USPQ2d 1329, 1336 (Fed. Cir. 2006)).

While the KSR Court rejected a rigid application of the teaching, suggestion, or motivation (“TSM”) test in an obviousness inquiry, the Court acknowledged the importance of identifying “a reason that would have prompted a person of ordinary skill in the relevant field to combine the elements in the way the claimed new invention does” in an obviousness determination. KSR, 127 S. Ct. at 1731. The court stated in dicta that, where there is a

“market pressure to solve a problem and there are a finite number of identified, predictable solutions, a person of ordinary skill has good reason to pursue the known options within his or her technical grasp. If this leads to the anticipated success, it is likely the product not of innovation but of ordinary skill and common sense. In that instance the fact that a combination was obvious to try might show that it was obvious under § 103.”

In a post-KSR decision, *PharmaStem Therapeutics, Inc. v. ViaCell, Inc.*, 491 F.3d 1342 (Fed. Cir. 2007), the Federal Circuit stated that:

an invention would not be invalid for obviousness if the inventor would have been motivated to vary all parameters or try each of numerous possible choices until one possibly arrived at a successful result, where the prior art gave either no indication of which parameters were critical or no direction as to which of many possible choices is likely to be successful. Likewise, an invention would not be deemed obvious if all that was suggested was to explore a new technology or general approach that seemed to be a promising field of experimentation, where the prior art gave only general guidance as to the particular form of the claimed invention or how to achieve it.

Furthermore, KSR has not overruled existing case law. See *In re Papesch*, (315 F.2d 381, 137 USPQ 43 (CCPA 1963)), *In re Dillon*, 919 F.2d 688, 16 USPQ2d 1897 (Fed. Cir. 1991), and *In re Deuel* (51 F.3d 1552, 1558-59, 34 USPQ2d 1210, 1215 (Fed. Cir. 1995)). “In cases involving new compounds, it remains necessary to identify some reason that would have led a chemist to modify a known compound in a particular manner to establish *prima facie* obviousness of a new claimed compound.” *Takeda v. Alphapharm*, 492 F.3d 1350 (Fed. Cir. 2007).

The mere fact that prior art may be modified to produce what is claimed does not make the modification obvious unless the prior art suggests the desirability of the modification. *In re Fritch*, 23 U.S.P.Q.2d 1780 (Fed. Cir. 1992); see, also, *In re Papesch*, 315 F.2d 381, 137 U.S.P.Q. 43 (CCPA 1963). In addition, if the proposed modification or combination of the prior art would change the principle of operation of the prior art invention being modified, then the teachings of the references are not sufficient to render the claims *prima facie* obvious. *In re Ratti*, 270 F.2d 810, 123 USPQ 349 (CCPA 1959).

The disclosure of the applicant cannot be used to hunt through the prior art for the claimed elements and then combine them as claimed. In *re Laskowski*, 871 F.2d 115, 117, 10 USPQ2d 1397, 1398 (Fed. Cir. 1989). “To imbue one of ordinary skill in the art with knowledge of the invention in suit, when no prior art reference or references of record convey or suggest that knowledge, is to fall victim to the insidious effect of a hindsight syndrome wherein that which only the inventor taught is used against its teacher” *W.L. Gore & Associates, Inc. v. Garlock Inc.*, 721 F.2d 1540, 1553, 220 USPQ 303, 312-13 (Fed. Cir. 1983).

B. THE REJECTION OF CLAIMS 1, 11-13, 34-36, 40-42, 113 AND 113 UNDER 35 U.S.C. §103(b) SHOULD BE REVERSED BECAUSE THE EXAMINER HAS FAILED TO ESTABLISH A PRIMA FACIE CASE OF OBVIOUSNESS

1. The teachings of O’Brien

The teachings of O’Brien are discussed above. O’Brien states that:

TADG-15 is a highly overexpressed gene in tumors. It is expressed in a limited number of normal tissues, primarily tissues that are involved in either uptake or secretion of molecules e.g. colon and pancreas. TADG-15 is further novel in its component structure of domains in that it has a protease catalytic domain which could be released and used as a diagnostic and which has the potential for a target for therapeutic intervention.

O’Brien is speculating that the protease domain could be released *in vivo* and serve as a therapeutic target **not as a therapeutic agent or diagnostic reagent**. O’Brien does not teach or suggest that the protease domain exists even *in vivo* as a single chain, and does not teach or suggest isolating it. In this passage, noted by the Examiner, O’Brien is discussing the expression of TADG-15 in tumors and other tissues and indicates that it is expressed on the surface of cells. Because of its structure, the protease domain could be presented on the surface of cells *in vivo*, and, thus, “could be released.” Since it is over expressed in tumors, if released *in vivo*, it could serve as a diagnostic marker indicating the presence of tumor cells. Use of its presence *in vivo* as a diagnostic marker for detection of tumors and/or as a therapeutic target is not a teaching or suggestion or hint for isolating the protease domain, nor for producing it as a single-chain polypeptide, nor for modifying it by replacing what would be a free Cys in a single chain form with another amino acid.

Thus, O’Brien does not state or hint that the isolated single chain protease domain could be used as therapeutic or as a diagnostic, and certainly does not teach or suggest then modifying it by replacing a free Cys in the single chain polypeptide with another amino acid. Such teaching does not constitute even a hint or suggestion for isolation or production of a polypeptide consisting only of the single-chain protease domain of an MTSP, nor of a single

chain protease domain in which the free Cys (which results only by virtue of it being a single chain) is replaced with another amino acid.

2. Analysis - the Examiner has failed to set forth a case of prima facie obviousness.

Independent Claim 1

O'Brien does not teach or suggest an isolated single chain protease domain of an MTSP polypeptide nor one in which a free Cys residue is replaced with another amino acid, such as a serine. There is no teaching or suggestion in O'Brien for preparing a polypeptide consisting only of a single-chain protease domain and modifying by replacing what is a free Cys in the single-chain form with another amino acid. The Examiner acknowledges that O'Brien does not teach a protease domain of an MTSP polypeptide where a free Cys residue in the protease domain is replaced with Ser residues. See, for example, the non-final Office Action, mailed June 25, 2007 (Exhibit 1), at page 25, which recites:

The reference O'Brien *et al.* does not teach a serine protease domain of a MTPSP [sic] polypeptides wherein free Cys residues have been replaced with Ser residues.

Even post-KSR, "it remains necessary to identify some reason that would have led a chemist to modify a known compound in a particular manner to establish prima facie obviousness of a new claimed compound." *Takeda Chem. Indus., Ltd. v. Alphapharm Pty., Ltd.* (Fed. Cir. 2007).

In this instance, there is no teaching or suggestion in O'Brien for isolating a single chain polypeptide consisting only of an MTSP protease domain in which a free Cys is replaced with another amino acid. O'Brien provides no teaching or suggestion for isolating the protease domain and preparing it as a single chain. O'Brien does not teach or suggest replacing any amino acid in the MTSP polypeptide with another amino acid, and provides no teaching or suggestion for modifying a single-chain polypeptide having a free Cys residue by replacing the free Cys residue with another amino acid.

For at least the reasons discussed above, O'Brien, alone or in combination with what was known in the art, does not teach or suggest every element of independent claim 1. Accordingly, Appellant respectfully submits that claim 1 is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 1. Appellant respectfully submits that the rejection of claim 1 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

For the reasons above, O'Brien fails to set forth a prima facie case of obvious of any of the dependent claims and further, additional reasons why O'Brien fails to set forth a prima facie case of obvious of each dependent claim are described below.

Dependent Claim 11

Claim 11 depends from claim 1 and specifies that the MTSP is selected from among MTSP1, MTSP3, MTSP4 and MTSP6. Claim 11 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien, alone or in combination with what was known in the art, does not teach or suggest every element of claim 11. Accordingly, Appellant respectfully submits that claim 11 is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 11. Appellant respectfully submits that the rejection of claim 11 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 12

Claim 12 depends from claim 1 and specifies that the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2 (MTSP1), amino acids 205-437 of SEQ ID NO. 4 (MTSP3), the amino acid residues set forth as SEQ ID No. 6 (MTSP4) or as amino acids 217-443 in SEQ ID No. 12 (MTSP6), where the free Cys is replaced with another amino acid. Claim 12 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien, alone or in combination with what was known in the art, does not teach or suggest every element of claim 12. Accordingly, Appellant respectfully submits that claim 12 is not taught or suggested by O'Brien.. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 12. Appellant respectfully submits that the rejection of claim 12 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 13

Claim 13 depends from claim 1 and specifies that the substantially purified polypeptide has at least about 95% sequence identity with a protease domain consisting of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acids set forth as SEQ ID No. 6, and amino acids 217-443 in SEQ ID No. 12. Claim 13 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien, alone or in combination with what was known in the art, does not teach or suggest every element of claim 13. Hence,

Appellant respectfully submits that claim is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 13. Appellant respectfully submits that the rejection of claim 13 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 34

Claim 34 depends from claim 1 and specifies that the MTSP is selected from among corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4. Claim 34 includes every limitation of claim 1, from which it depends. Thus, for the reasons discussed above with respect to claim 1, O'Brien, alone or in combination with what was known in the art, does not teach or suggest every element of claim 34. Accordingly, Appellant respectfully submits that claim 34 is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 34. Appellant respectfully submits that the rejection of claim 34 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 35

Claim 35 is directed to a conjugate that comprises a) a polypeptide of claim 1 and b) a targeting agent linked to the protein directly or via a linker, wherein the conjugate has serine protease activity. The specification defines a targeting agent as

any moiety, such as a protein or effective portion thereof, that provides specific binding of the conjugate to a cell surface receptor, which, preferably, internalizes the conjugate or MTSP portion thereof. A targeting agent may also be one that promotes or facilitates, for example, affinity isolation or purification of the conjugate; attachment of the conjugate to a surface; or detection of the conjugate or complexes containing the conjugate.

(*e.g.*, see page 38, lines 9-15).

Claim 35 recites that a targeting agent is linked to the protein of claim 1 directly or via a linker and that the conjugate has serine protease activity. There is no teaching or suggestion in O'Brien of conjugating a targeting agent to an isolated single-chain polypeptide consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid.

O'Brien teaches, at col. 9, lines 53-56, covalently linking another polypeptide to an intact TAGD-15 polypeptide or to a fragment thereof. The cited section states:

The fragment, or the intact TAGD-15 polypeptide, may be covalently linked to another polypeptide, *e.g.*, which acts as a label, a ligand, or a means to increase **antigenicity**. [emphasis added]

By “fragment” O’Brien mean “antigenic fragment” or other fragment (see, col. 9, lines, 22-32), which describe fragments as 10 residues, typically 20 residues and “preferably at least 30 (e.g 50) residues” in length, and indicates that they can be antigenic fragments for preparing antibodies. From the context, O’Brien contemplates antigenic fragments. There is no mention, teaching suggestion or hint that the fragment is a catalytic domain or fragment thereof. .

O’Brien does not teach or suggest isolating the protease domain of TADG-15 and conjugating it to another polypeptide. The Examiner alleges that the motivation for making conjugates is to use it as a diagnostic, which has the potential for a target for therapeutic intervention (page 23 of the Office Action). Even if there were such suggestion in O’Brien, as noted above, there is no teaching or suggestion for isolating the protease domain or a catalytically active portion thereof and replacing a free Cys residue. Hence there can be no motivation to prepare conjugates. Furthermore, as discussed above, O’Brien suggests isolating antigenic fragments, and linking them to another polypeptide, such as a label, ligand or as means to increase antigenicity. O’Brien contemplates using antigenic fragments to make antibodies because the TAGD-15 polypeptide is considered a possible therapeutic target, not as a therapeutic agent or as a diagnostic agent.

O’Brien teaches that TADG-15 is a highly over-expressed gene in tumors and suggests that TADG-15 thus could be a potential target for therapeutic intervention (col. 15, lines 31-38). One of ordinary skill in the art would not be lead to conjugate a targeting moiety to a target. O’Brien does not teach, suggest or mention conjugating a targeting agent to an isolated protease domain. Accordingly, for these reasons and the reasons discussed above with respect to claim 1, Appellant respectfully submits that claim 35 is not taught or suggested by O’Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 35. Appellant respectfully submits that the rejection of claim 35 as obvious over O’Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 36

Claim 36 depends from claim 35 and recites that the targeting agent permits i) affinity isolation or purification of the conjugate; ii) attachment of the conjugate to a surface; iii) detection of the conjugate; or iv) targeted delivery to a selected tissue or cell. As discussed above, O’Brien does not teach or suggest isolating the protease domain of TADG-15, replacing a free Cys with another amino acid and conjugating the single chain protease

domain to a targeting agent. Accordingly, for these reasons and the reasons discussed above with respect to claim 1, Appellant respectfully submits that claim 36 not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 36. Appellant respectfully submits that the rejection of claim 36 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 40

Claim 40 is directed to a solid support comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker. O'Brien does not mention a solid support. There is no teaching or suggestion in O'Brien of a solid support that includes two or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. In maintaining the rejection, the Examiner states that "assays using polypeptides linked to the molecules taught by O'Brien *et al.* utilize solid supports" (page 23 of the Office Action). In the assays described in O'Brien, a hybridization probe to the nucleotide encoding TAGD-15 polypeptide (such as in a standard Northern blot assay) or an antibody to the TAGD-15 polypeptide (such as in a standard immunoassay) is attached to a solid support. Appellant respectfully submits that, although such assays can use solid supports, O'Brien does not teach or suggest an isolated single-chained polypeptide consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid nor conjugating two or more such isolated protease domains to a solid support. Accordingly, for these reasons and the reasons discussed above with respect to claim 1, Appellant respectfully submits that claim 40 is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 40. Appellant respectfully submits that the rejection of claim 40 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 41

Claim 41 recites a solid support comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker where the polypeptides comprise an array. The specification defines an array as a collection of elements containing three or more members. As discussed above, O'Brien does not mention a solid support. O'Brien provides no teaching or suggestion for isolating the protease domain and preparing it as a single chain. There is no teaching or suggestion in O'Brien of a solid support that includes three or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was

replaced with another amino acid. Accordingly, for these reasons and the reasons discussed above with respect to claim 1, claim 41 is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 41. Appellant respectfully submits that the rejection of claim 41 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 42

Claim 42 is directed to the solid support of claim 41, wherein the array comprises polypeptides having different MTSP protease domains. There is no teaching or suggestion in O'Brien of a solid support that includes three or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Further, the only MTSP taught in O'Brien is TAGD-15. There is no teaching or suggestion of any other MTSP. Hence, there can be no teaching or suggestion in O'Brien to conjugate isolated protease domains from different MTSPs to a solid support to form an array. Accordingly, for these reasons and the reasons discussed above with respect to claim 1, Appellant respectfully submits that claim 42 is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 42. Appellant respectfully submits that the rejection of claim 42 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 113

Claim 113 is directed to a solid support comprising two or more polypeptides of claim 12 linked thereto either directly or via a linker. Claim 12 depends from claim 1 and recites that the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12. Claim 12 includes every limitation of claim 1, from which it depends.

O'Brien does not mention a solid support. Furthermore, there is no teaching or suggestion in O'Brien of a solid support that includes two or more isolated single-chain polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Accordingly, for these reasons and the reasons discussed above with respect to claim 1, Appellant respectfully submits that claim 113 is not taught or suggested by O'Brien the Examiner has failed to set forth a prima facie case of obviousness

of claim 113. Appellant respectfully submits that the rejection of claim 113 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Dependent Claim 114

Claim 114 depends from claim 113 and is directed to an array. The specification defines an array as a collection of elements containing three or more members. O'Brien provides no teaching or suggestion of an array that includes three or more isolated single-chained polypeptides consisting only of an MTSP protease domain in which a free Cys was replaced with another amino acid. Accordingly, for these reasons and the reasons discussed above with respect to claim 1, claim 114 is not taught or suggested by O'Brien. Thus, the Examiner has failed to set forth a prima facie case of obviousness of claim 114. Appellant respectfully submits that the rejection of claim 114 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

Summary

Appellant respectfully submits that claim 1 as well as each of claims 11-13, 34-36, 40-42, 113 and 114, which ultimately depend from claim 1 and include every limitation thereof, are nonobvious and distinguishable from the teachings of O'Brien. Thus, Appellant respectfully submits that the Examiner has failed to establish claims 1, 11-13, 34-36, 40-42, 113 and 114 as obvious under 35 U.S.C. §103(a) over O'Brien. Accordingly, Appellant respectfully submits that the rejection of claims 1, 11-13, 34-36, 40-42, 113 and 114 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.

VIII. CONCLUSIONS

Appellant respectfully submits that the rejection of claims 1, 11, 20, 34-36, 40-42, 113 and 114 under 35 U.S.C. §112, first paragraph, as allegedly containing subject matter that was not described in the specification in such a way as to reasonably convey to one skilled in the art that the inventor, at the time the application was filed, had possession of the claimed subject matter, is erroneous in law and fact and, therefore, should be reversed.

Appellant also respectfully submits that the rejection of claims 1, 11, 20, 34-36, 40-42, 113 and 114 under 35 U.S.C. § 112, first paragraph, because the specification allegedly fails to describe the claimed subject matter in such a way as to enable one skilled in the art to make and use the claimed subject matter commensurate in scope with these claims, is erroneous in law and fact and, therefore, should be reversed.

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

Appellant also respectfully submits that the Examiner has failed to establish claims 1, 11-13, 20, 34-36, 40-42, 113 and 114 as anticipated by Takeuchi under 35 U.S.C. §102(b). Accordingly, Appellant respectfully submits that the rejection of claims 1-3, 19 and 20 as anticipated by Takeuchi is erroneous in law and fact and, therefore, should be reversed.

Appellant also respectfully submits that the Examiner has failed to establish claims 1, 11-13 and 34 as anticipated by O'Brien under 35 U.S.C. §102(e). Accordingly, Appellant respectfully submits that the rejection of claims 1, 11-13 and 34 as anticipated by O'Brien is erroneous in law and fact and, therefore, should be reversed.

Appellant further respectfully submits that the Examiner has failed to establish claims 1, 11-13, 34-36, 40-42, 113 and 114 as obvious under 35 U.S.C. §103(a) over O'Brien. Accordingly, Appellant respectfully submits that the rejection of claims 1, 11-13, 34-36, 40-42, 113 and 114 as obvious over O'Brien is erroneous in law and fact and, therefore, should be reversed.


* * *

The Director is authorized to charge any fees that may be required, or to credit any overpayment to Deposit Account No. 02-1818. Please indicate the Attorney Docket No. 119385-00028/1607 on the account statement. If a Petition for Extension of Time is needed, this paper is to be considered such Petition.

Respectfully submitted,

Dated: March 16, 2009

BY:


Stephanie Seidman
Reg. No. 33,779

Address all correspondence to:
77202
Stephanie Seidman
K&L Gates LLP
3580 Carmel Mountain Road, Suite 200
San Diego, California 92130
Telephone: (858) 509-7410
Facsimile: (858) 509-7460
email: stephanie.seidman@klgates.com

CLAIMS APPENDIX

PENDING CLAIMS ON APPEAL OF U.S. PATENT APPLICATION SERIAL NO. 09/776,191

1. (Rejected) An isolated, substantially purified single-chain poly-peptide, consisting only of a protease domain of a type-II membrane-type serine protease (MTSP) or a catalytically active fragment thereof as a single chain, wherein:

a free Cys in the protease domain is replaced with another amino acid; and
the MTSP protease domain or catalytically active fragment thereof has serine protease activity as a single chain.

2. - 9. (Cancelled).

10. (Withdrawn) The substantially purified polypeptide of claim 1, wherein the MTSP portion has an N-terminus that comprises IVNG, ILGG, VGLL or ILGG.

11. (Rejected) The substantially purified polypeptide of claim 1, wherein the MTSP is selected from among MTSP1, MTSP3, MTSP4 and MTSP6.

12. (Rejected) The substantially purified polypeptide of claim 1, wherein the MTSP protease domain consists of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acid residues set forth as SEQ ID No. 6 or as amino acids 217-443 in SEQ ID No. 12.

13. (Rejected) The substantially purified polypeptide of claim 1 that has at least about 95% sequence identity with a protease domain consisting of a sequence of amino acid residues selected from among amino acids 615-855 of SEQ ID No. 2, amino acids 205-437 of SEQ ID NO. 4, the amino acids set forth as SEQ ID No. 6, and amino acids 217-443 in SEQ ID No. 12.

Claims 14 - 19 (Cancelled).

20. (Rejected) The polypeptide of claim 1, wherein a free Cys in the protease domain is replaced with a serine.

Claims 21- 33 (Cancelled).

34. (Rejected) The polypeptide of claim 1, wherein the MTSP is selected from among corin, MTSP1, enteropeptidase, human airway trypsin-like protease (HAT), TMPRSS2, and TMPRSS4.

35. (Rejected) A conjugate, comprising:

- a) a polypeptide of claim 1, and
- b) a targeting agent linked to the protein directly or via a linker, wherein the conjugate has serine protease activity.

36. (Rejected) The conjugate of claim 35, wherein the targeting agent permits

- i) affinity isolation or purification of the conjugate;
- ii) attachment of the conjugate to a surface;
- iii) detection of the conjugate; or
- iv) targeted delivery to a selected tissue or cell.

Claims 37 – 39 (Cancelled)

40. (Rejected) A solid support comprising two or more polypeptides of claim 1 linked thereto either directly or via a linker.

41. (Rejected) The support of claim 40, wherein the polypeptides comprise an array.

42. (Rejected) The support of claim 41, wherein the array comprises polypeptides having different MTSP protease domains.

43. (Withdrawn) A method for identifying candidate anti-tumor compounds that inhibit the protease activity of an MTSP, comprising:

contacting a polypeptide of claim 1 with a substrate proteolytically cleaved by the MTSP, and, either simultaneously, before or after, adding a test compound or plurality thereof; measuring the amount of substrate cleaved in the presence of the test compound; and selecting compounds that change the amount cleaved compared to a control, whereby compounds that modulate the activity of the MTSP are identified.

44. (Withdrawn) The method of claim 43, wherein the test compounds are small molecules, peptides, peptidomimetics, natural products, antibodies or fragments thereof.

45. (Withdrawn) The method of claim 43, wherein a plurality of the test compounds are screened simultaneously.

46. (Withdrawn) The method of claim 43, wherein the change in the amount cleaved is assessed by comparing the amount cleaved in the presence of the test compound with the amount in the absence of the test compound.

47. (Cancelled)

48. (Withdrawn) The method of claim 43, wherein a plurality of the polypeptides are linked to a solid support, either directly or via a linker.

49. (Withdrawn) The method of claim 43, wherein the polypeptides comprise an array.

50. (Withdrawn) The method of claim 43, wherein the polypeptides comprise a plurality of different MTSP proteases.

51. (Withdrawn) A method of identifying a compound that specifically binds to a single chain protease domain of an MTSP, comprising:

contacting a polypeptide of claim 1 with a test compound or plurality thereof under conditions conducive to binding thereof; and

identifying compounds that specifically bind to the MTSP single chain protease domain or compounds that inhibit binding of a compound known to bind to the MTSP single chain protease domain, wherein the known compound is contacted with the polypeptide before, simultaneously with or after the test compound.

52. (Withdrawn) The method of claims 51, wherein the polypeptide is linked either directly or indirectly via a linker to a solid support.

53. (Withdrawn) The method of claim 51, wherein the test compounds are small molecules, peptides, peptidomimetics, natural products, antibodies or fragments thereof.

54. (Withdrawn) The method of claim 51, wherein a plurality of the test substances are screened for simultaneously.

55. (Withdrawn) The method of claim 52, wherein a plurality of the polypeptides are linked to a solid support.

56. -107. (Cancelled).

108. (Withdrawn) A conjugate, comprising:

- a) an MTSP3 or an MTSP4 or the MTSP6 of claim 12; and
- b) a targeting agent linked to the protein directly or via a linker.

109. (Withdrawn) The conjugate of claim 108, wherein the targeting agent permits

- i) affinity isolation or purification of the conjugate;
- ii) attachment of the conjugate to a surface;
- iii) detection of the conjugate; or
- iv) targeted delivery to a selected tissue or cell.

Claims 110 – 112 (Cancelled).

113. (Rejected) A solid support comprising two or more polypeptides of claim 12 linked thereto either directly or via a linker

114. (Rejected) The support of claim 113, wherein the polypeptides comprise an array.

115. (Withdrawn) A method for identifying compounds that modulate the protease activity of an MTSP of claim 1, comprising:

contacting the MTSP of claim 1 with a substrate proteolytically cleaved by the MTSP, and, either simultaneously, before or after, adding a test compound or plurality thereof; measuring the amount of substrate cleaved in the presence of the test compound; and selecting compounds that change the amount cleaved compared to a control, whereby compounds that modulate the activity of the MTSP are identified.

116. (Withdrawn) The method of claim 115, wherein the test compounds are small molecules, peptides, peptidomimetics, natural products, antibodies or fragments thereof.

117. (Cancelled).

118. (Withdrawn) The method of claim 115, wherein the change in the amount cleaved is assessed by comparing the amount cleaved in the presence of the test compound with the amount in the absence of the test compound.

119. (Withdrawn) The method of claim 115, wherein a plurality of the test substances are screened for simultaneously.

120. (Withdrawn) The method of claim 119, wherein a plurality of the polypeptides are linked to a solid support.

121. (Cancelled).

122. (Withdrawn) A method of identifying a compound that specifically binds to an MTSP protease domain, comprising:

contacting an MTSP protease domain of claim 12 with a test compound or plurality thereof under conditions conducive to binding thereof; and identifying compounds that specifically bind to the MTSP.

123. (Withdrawn) The method of claim 122, wherein the polypeptide is linked either directly or indirectly via a linker to a solid support.

124. (Withdrawn) The method of claim 122, wherein the test compounds are small molecules, peptides, peptidomimetics, natural products, antibodies or fragments thereof.

125. (Withdrawn) The method of claim 122, wherein a plurality of the test substances are screened for simultaneously.

126. (Withdrawn) The method of claim 125, wherein a plurality of the polypeptides are linked to a solid support.

127. – 137. (Cancelled).

EVIDENCE APPENDIX

- EXHIBIT 1: Final Office Action, dated March 26, 2008.
- EXHIBIT 2: Non-final Office Action, dated June 25, 2007.
- EXHIBIT 3: Takeuchi *et al.*, Proc. Natl. Acad. Sci. USA 96: 11054-11061 (1999).
- EXHIBIT 4: O'Brien *et al.*, U.S. Patent No. 5,972,616.
- EXHIBIT 5: Bachovchin *et al.*, Proc. Natl Acad. Sci. 78: 7323-7326 (1981).
- EXHIBIT 6: Brinkley, "A Brief Survey of Methods for Preparing Protein Conjugates with Dyes, Haptens, and Cross-linking Reagents" in Perspectives in Bioconjugate Chemistry (Claude Meares, ed. 1993, Chapter 4, pages 59-70).
- EXHIBIT 7: Bryan, Biochem. Biophys. Acta 1543: 200-203 (2000).
- EXHIBIT 8: Carter *et al.*, Nature 332: 564-568 (1988).
- EXHIBIT 9: Cheah *et al.*, J. Biol. Chem. 265: 7180-7187 (1990).
- EXHIBIT 10: Craik *et al.*, Science 237:909-913 (1987).
- EXHIBIT 11: Dawson *et al.*, U.S. Pat. No. 5,645,833 (1997).
- EXHIBIT 12: Devereux *et al.*, Nucleic Acids Research 12(I):387-395 (1984).
- EXHIBIT 13: Farley *et al.*, Biochem. Biophys. Acta 1173: 350-352 (1993).
- EXHIBIT 14: Hooper *et al.*, Eur. J. Biochem. 267: 6931-6937 (2000).
- EXHIBIT 15: Hooper *et al.*, J. Biol. Chem. 276: 857-860 (2001).
- EXHIBIT 16: Jacquinet *et al.*, FEBS Lett. 468: 93-100 (2000).
- EXHIBIT 17: Kitamoto *et al.*, Proc Natl Acad Sci USA 91: 7588-7592 (1994).
- EXHIBIT 18: Kitamoto *et al.*, Biochem. 27: 4562-4568 (1995).
- EXHIBIT 19: Leytus *et al.*, Biochemistry 27: 1067-1074 (1988).
- EXHIBIT 20: Lin *et al.*, J. Biol. Chem. 274: 18231-18236 (1999).
- EXHIBIT 21: Lu *et al.*, J. Mol. Biol. 292: 361-373 (1999).
- EXHIBIT 22: Matsushima *et al.*, J. Biol. Chem. 269: 19976-19982 (1994).
- EXHIBIT 23: Means & Feeney, "Chemical Modifications of Proteins: History and Applications" in Perspectives in Bioconjugate Chemistry (Claude Meares, ed., 1993, Chapter 2, pages 10-20).
- EXHIBIT 24: Nienaber *et al.*, J. Biol. Chem. 275: 7239-48 (2000).
- EXHIBIT 25: O'Brien *et al.*, International PCT application No. WO 00/52044.
- EXHIBIT 26: Parks *et al.*, J. Biol. Chem. 268: 19101-19109 (1993).
- EXHIBIT 27: Parks & Lamb, Cell 64: 777-787 (1991).
- EXHIBIT 28: Pearson *et al.*, Proc. Natl. Acad. Sci. USA 85: 2444 (1988).
- EXHIBIT 29: Pearson *et al.*, Cabios Invited Review 13(4): 325-332 (1997).
- EXHIBIT 30: Perona & Craik, Protein Science 4: 337-360 (1995).

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

- EXHIBIT 31: Paoloni-Giacobino *et al.*, Genomics 44: 309-320 (1997).
EXHIBIT 32: Silverman *et al.*, Curr. Opin. Chem. Biol., 2: 397-403 (1998).
EXHIBIT 33: Sittampalam *et al.*, Curr. Opin. Chem. Biol., 1: 384-391 (1997).
EXHIBIT 34: Sommerhoff *et al.*, Proc. Natl. Acad. Sci. USA 96:10984-10991 (1999).
EXHIBIT 35: Sprang *et al.*, Science 237: 905-909 (1987).
EXHIBIT 36: Tomita *et al.*, J. Biochem. 124: 784-789 (1998).
EXHIBIT 37: Tsuji *et al.*, J Biol Chem 266(25): 16948-16953 (1991).
EXHIBIT 38: Vu *et al.*, J. Biol. Chem. 272: 31315-31320 (1997).
EXHIBIT 39: Wallrapp *et al.*, Cancer 60: 2602-2606 (2000).
EXHIBIT 40: Walter *et al.*, Annu. Rev. Cell Biol. 2: 499-516 (1986).
EXHIBIT 41: Xu *et al.*, J. Biol. Chem. 275: 378-385 (2000).
EXHIBIT 42: Yahagi *et al.*, Biochem. Biophys. Res. Commun. 219: 806-812 (1996).
EXHIBIT 43: Yamaoka *et al.*, J. Biol. Chem. 273: 11895-11901 (1998).
EXHIBIT 44: Yan *et al.*, J. Biol. Chem. 274: 14926-14935 (1999).
EXHIBIT 45: Zubay, Biochemistry (1983), pages 12-13.
EXHIBIT 46: Office Action, mailed April 21, 2006.

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Document No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

RELATED PROCEEDINGS APPENDIX

None

EVIDENCE APPENDIX

- EXHIBIT 1: Final Office Action, dated March 26, 2008.
- EXHIBIT 2: Non-final Office Action, dated June 25, 2007.
- EXHIBIT 3: Takeuchi *et al.*, Proc. Natl. Acad. Sci. USA 96: 11054-11061 (1999).
- EXHIBIT 4: O'Brien *et al.*, U.S. Patent No. 5,972,616.
- EXHIBIT 5: Bachovchin *et al.*, Proc. Natl Acad. Sci. 78: 7323-7326 (1981).
- EXHIBIT 6: Brinkley, "A Brief Survey of Methods for Preparing Protein Conjugates with Dyes, Haptens, and Cross-linking Reagents" in Perspectives in Bioconjugate Chemistry (Claude Meares, ed. 1993, Chapter 4, pages 59-70).
- EXHIBIT 7: Bryan, Biochem. Biophys. Acta 1543: 200-203 (2000).
- EXHIBIT 8: Carter *et al.*, Nature 332: 564-568 (1988).
- EXHIBIT 9: Cheah *et al.*, J. Biol. Chem. 265: 7180-7187 (1990).
- EXHIBIT 10: Craik *et al.*, Science 237:909-913 (1987).
- EXHIBIT 11: Dawson *et al.*, U.S. Pat. No. 5,645,833 (1997).
- EXHIBIT 12: Devereux *et al.*, Nucleic Acids Research 12(I):387-395 (1984).
- EXHIBIT 13: Farley *et al.*, Biochem. Biophys. Acta 1173: 350-352 (1993).
- EXHIBIT 14: Hooper *et al.*, Eur. J. Biochem. 267: 6931-6937 (2000).
- EXHIBIT 15: Hooper *et al.*, J. Biol. Chem. 276: 857-860 (2001).
- EXHIBIT 16: Jacquinet *et al.*, FEBS Lett. 468: 93-100 (2000).
- EXHIBIT 17: Kitamoto *et al.*, Proc Natl Acad Sci USA 91: 7588-7592 (1994).
- EXHIBIT 18: Kitamoto *et al.*, Biochem. 27: 4562-4568 (1995).
- EXHIBIT 19: Leytus *et al.*, Biochemistry 27: 1067-1074 (1988).
- EXHIBIT 20: Lin *et al.*, J. Biol. Chem. 274: 18231-18236 (1999).
- EXHIBIT 21: Lu *et al.*, J. Mol. Biol. 292: 361-373 (1999).
- EXHIBIT 22: Matsushima *et al.*, J. Biol. Chem. 269: 19976-19982 (1994).
- EXHIBIT 23: Means & Feeney, "Chemical Modifications of Proteins: History and Applications" in Perspectives in Bioconjugate Chemistry (Claude Meares, ed., 1993, Chapter 2, pages 10-20).
- EXHIBIT 24: Nienaber *et al.*, J. Biol. Chem. 275: 7239-48 (2000).
- EXHIBIT 25: O'Brien *et al.*, International PCT application No. WO 00/52044.
- EXHIBIT 26: Parks *et al.*, J. Biol. Chem. 268: 19101-19109 (1993).
- EXHIBIT 27: Parks & Lamb, Cell 64: 777-787 (1991).
- EXHIBIT 28: Pearson *et al.*, Proc. Natl. Acad. Sci. USA 85: 2444 (1988).
- EXHIBIT 29: Pearson *et al.*, Cabios Invited Review 13(4): 325-332 (1997).
- EXHIBIT 30: Perona & Craik, Protein Science 4: 337-360 (1995).

Applicant : Madison *et al.*
Serial No. : 09/776,191
Filed : February 2, 2001
Customer Number: 77202

Attorney's Docket No.: 119385-00028 / 1607
APPELLANT'S APPEAL BRIEF

- EXHIBIT 31: Paoloni-Giacobino *et al.*, Genomics 44: 309-320 (1997).
EXHIBIT 32: Silverman *et al.*, Curr. Opin. Chem. Biol., 2: 397-403 (1998).
EXHIBIT 33: Sittampalam *et al.*, Curr. Opin. Chem. Biol., 1: 384-391 (1997).
EXHIBIT 34: Sommerhoff *et al.*, Proc. Natl. Acad. Sci. USA 96:10984-10991 (1999).
EXHIBIT 35: Sprang *et al.*, Science 237: 905-909 (1987).
EXHIBIT 36: Tomita *et al.*, J. Biochem. 124: 784-789 (1998).
EXHIBIT 37: Tsuji *et al.*, J Biol Chem 266(25): 16948-16953 (1991).
EXHIBIT 38: Vu *et al.*, J. Biol. Chem. 272: 31315-31320 (1997).
EXHIBIT 39: Wallrapp *et al.*, Cancer 60: 2602-2606 (2000).
EXHIBIT 40: Walter *et al.*, Annu. Rev. Cell Biol. 2: 499-516 (1986).
EXHIBIT 41: Xu *et al.*, J. Biol. Chem. 275: 378-385 (2000).
EXHIBIT 42: Yahagi *et al.*, Biochem. Biophys. Res. Commun. 219: 806-812 (1996).
EXHIBIT 43: Yamaoka *et al.*, J. Biol. Chem. 273: 11895-11901 (1998).
EXHIBIT 44: Yan *et al.*, J. Biol. Chem. 274: 14926-14935 (1999).
EXHIBIT 45: Zubay, Biochemistry (1983), pages 12-13.
EXHIBIT 46: Office Action, mailed April 21, 2006.

Exhibit 1



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
-----------------	-------------	----------------------	---------------------	------------------

09/776,191

02/02/2001

Edwin L. Madison

119385-00028 / 1607

3237

20985 7590 03/26/2008
FISH & RICHARDSON, PC
P.O. BOX 1022
MINNEAPOLIS, MN 55440-1022

EXAMINER

PAK, YONG D

ART UNIT

PAPER NUMBER

1652

MAIL DATE

DELIVERY MODE

03/26/2008

PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary

Application No.

09/776,191

Applicant(s)

MADISON ET AL.

Examiner

Yong D. Pak

Art Unit

1652

– The MAILING DATE of this communication appears on the cover sheet with the correspondence address –

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 26 December 2007.
- 2a) ☒ This action is **FINAL**. 2b) ☐ This action is non-final.
- 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) See Continuation Sheet is/are pending in the application.
- 4a) Of the above claim(s) 10,43-46,48-55,108,109,115,116,118-120 and 122-126 is/are withdrawn from consideration.
- 5) ☐ Claim(s) _____ is/are allowed.
- 6) ☒ Claim(s) 1,11-13,20,34-36,40-42,113 and 114 is/are rejected.
- 7) ☐ Claim(s) _____ is/are objected to.
- 8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☐ The specification is objected to by the Examiner.
- 10) ☐ The drawing(s) filed on _____ is/are: a) ☐ accepted or b) ☐ objected to by the Examiner.
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
- a) ☐ All b) ☐ Some * c) ☐ None of:
1. ☐ Certified copies of the priority documents have been received.
 2. ☐ Certified copies of the priority documents have been received in Application No. _____.
 3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).
- * See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- | | |
|---|---|
| 1) <input type="checkbox"/> Notice of References Cited (PTO-892) | 4) <input type="checkbox"/> Interview Summary (PTO-413)
Paper No(s)/Mail Date: _____ |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948) | 5) <input type="checkbox"/> Notice of Informal Patent Application |
| 3) <input checked="" type="checkbox"/> Information Disclosure Statement(s) (PTO/SB/08)
Paper No(s)/Mail Date <u>12/26/07</u> . | 6) <input type="checkbox"/> Other: _____ |

Continuation of Disposition of Claims: Claims pending in the application are 1,10-13,20,34-36,40-46,48-55,108,109,113-116,118-120 and 122-126.

DETAILED ACTION

This application is a CIP of 09/657,986, now issued as U.S. Patent No. 6,797,504.

The amendment filed on December 26, 2007, amending claim 1 and canceling claims 2-3 and 19, has been entered.

Claims 1, 10-13, 20, 34-36, 40-46, 48-55, 108-109, 113-116, 118-120 and 122-126 are pending. Claims 10, 43-46, 48-55, 108-109, 115-116, 118-120 and 122-126 are withdrawn. Claims 1, 11-13, 20, 34-36, 40-42 and 113-114 are under consideration.

Priority

Applicant's claim for domestic priority under 35 U.S.C. 119(e) is acknowledged. However, the provisional applications upon which priority is claimed fails to provide adequate support under 35 U.S.C. 112 for claims 11-13 and 34 of this application.

Provisional applications 60/179,982, 60/183,542, 60/213,124, 60/220,970 and 60/234,840 fail to provide adequate support for polypeptides comprising the serine protease domain of MTSP1. Provisional applications 60/179,982 and 60/183,542 describe polypeptides related MTSP3 and provisional application 60/213,124, 60/220,970 and 60/234,840 describe polypeptides related to MTSP4.

Therefore, the effective filing date for purpose of prior art is the filing date of 09/657,986, which is 9/8/2000.

Information Disclosure Statement

The information disclosure statement (IDS) submitted on December 26, 2007 was filed after the mailing date of the Non-Final Rejection on June 25, 2007. The submission is in compliance with the provisions of 37 CFR 1.97. Accordingly, the information disclosure statement is being considered by the examiner.

Response to Arguments

Applicant's amendment and arguments filed on December 26, 2007, have been fully considered and are deemed to be persuasive to overcome some of the rejections previously applied. Rejections and/or objections not reiterated from previous office actions are hereby withdrawn.

Claim Objections

Applicants argue that claims 11-13 and 34 should be retained pending a determination of the allowability of claim 1, which is a linking claim, linking the elected subject matter. In view of applicant's argument, the objection to claims 11-13 and 34 have been **withdrawn**.

Claim Rejections - 35 USC § 112 – 2nd paragraph

In view of applicant's argument, the rejection of claims 1, 11-13 and claims 20, 34-36, 40-42 and 113-114 depending therefrom under 35 U.S.C. 112, second

paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention has been **withdrawn**.

Claim Rejections - 35 USC § 112 – 1st paragraph

The following is a quotation of the first paragraph of 35 U.S.C. 112:

The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

Claims 1, 11, 20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

Claims 1, 11, 20, 34-36, 40-42 and 113-114 are drawn to a polypeptide consisting of a protease domain or catalytically active fragment thereof of type-II membrane-type serine protease (MTSP) from any source. Claims 11 and 34 limit the MTSP polypeptide to a MTSP1 polypeptide from any source. Therefore, these claims are drawn to a genus of polypeptides having any structure. The specification only teaches four species, amino acids 615-855 of SEQ ID NO:2 (MTSP1), amino acids of 205-437 of SEQ ID NO:4 (MTSP3), amino acids of SEQ ID NO:6 (MTSP4) and amino acids 217-443 of SEQ ID NO:11 (MTSP6). These species are not enough to describe the whole genus and there is no evidence on the record of the relationship between the

structure of the above catalytically active protease domains of SEQ ID NOs: 2, 4, 6 and 11 and the structure of the serine protease domain of any or all MTSP polypeptides or MTSP1 polypeptides. Further, the specification does not describe the structure of a catalytically active fragment of a protease domain of any or all MTSP polypeptide. Therefore, the specification fails to describe a representative species of the genus of polypeptides consisting of a serine protease domain or a catalytically active portion of a MTSP polypeptide.

Given this lack of description of the representative species encompassed by the genus of the claims, the specification fails to sufficiently describe the claimed invention in such full, clear, concise, and exact terms that a skilled artisan would recognize that applicants were in possession of the inventions of claims 1, 11, 20, 34-36, 40-42 and 113-114.

Applicant is referred to the revised guidelines concerning compliance with the written description requirement of U.S.C. 112, first paragraph, published in the Official Gazette and also available at www.uspto.gov.

In response to the previous Office Action, applicants have traversed the above rejection.

Applicants argue that the claims are fully described because the specification identified 17 members of the MTSP family and identifies the protease domains thereof, unknown MTSPs and its protease domains. Examiner respectfully disagrees. The claims are not limited to specific protease domains of specific MTSP proteins, but the claims are drawn to polypeptides consisting of any protease domains or any or all

catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. The recitation of "protease domain of a MTSP" or "MTSP1" fails to provide a sufficient description of the claimed genus of polypeptides as it merely describes the functional features of the genus without providing any definition of the structural features of the species within the genus. The CAFC in *UC California v. Eli Lilly*, (43 USPQ2d 1398) stated that: "in claims to genetic material, however a generic statement such as 'vertebrate insulin cDNA' or 'mammalian insulin cDNA,' without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function. It does not specifically define any of the genes that fall within its definition. It does not define any structural features commonly possessed by members of the genus that distinguish them from others. One skilled in the art therefore cannot, as one can do with a fully described genus, visualize or recognize the identity of the members of the genus." Similarly with the claimed genus of protease domains, the functional definition of the genus does not provide any structural information commonly possessed by members of the genus which distinguish the species within the genus from other proteins such that one can visualize or recognize the identity of the members of the genus.

Further, as discussed in the written description guidelines, the written description requirement for a claimed genus may be satisfied through sufficient description of a representative number of species by actual reduction to practice, reduction to drawings, or by disclosure of relevant, identifying characteristics, i.e., structure or other physical

and/or chemical properties, by functional characteristics coupled with a known or disclosed correlation between function and structure, or by a combination of such identifying characteristics, sufficient to show the applicant was in possession of the claimed genus. A representative number of species means that the species which are adequately described are representative of the entire genus. **Thus, when there is substantial variation within the genus, one must describe a sufficient variety of species to reflect the variation within the genus.** Satisfactory disclosure of a representative number depends on whether one of skill in the art would recognize that the applicant was in possession of the necessary common attributes or features of the elements possessed by the members of the genus in view of the species disclosed. For inventions in an unpredictable art, adequate written description of a genus which embraces widely variant species cannot be achieved by disclosing only one species within the genus. In the instant case the claimed genera of the claims are drawn to species which are widely variant in structure. The genus of the claims are structurally diverse as it encompasses any catalytically active protease domains of any or all MTSP or MTSP1, excepting having serine protease activity. As such, neither the description of solely structural features present in all members of the genus is sufficient to be representative of the attributes and features of the entire genus.

Applicants also argue that the claims are fully described because members of the MTSP family of serine proteases were well known at the time of filing, such as conserved characteristic structural elements and protease domains and method of identifying serine protease domains were known in the art. Examiner respectfully

disagrees. As discussed above, the claims are not drawn to the specific protease domains of specific MTSP type II, but to polypeptides consisting of any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. In view of the widely variant species encompassed by the genus, the species disclosed in the specification is not enough and does not constitute a representative number of species to describe the whole genus of any or all variants, recombinant and mutants of any or all polypeptides having serine protease activity isolated from any or all source, including any or all variants, recombinants and mutants thereof, and there is no evidence on the record of the relationship between the structure of the protease domain of the specific MTSPs disclosed in the specification and the structure of any or all recombinant, variant and mutant of any or all polypeptides having serine protease activity. Therefore, the specification fails to describe a representative species of the genus comprising any or all polypeptides having serine protease activity, including any or all variants, recombinants and mutants thereof.

Applicants also argue that the claims are fully described by the specification because one skilled in the art would recognize applicant's possession of the claimed subject matter. Examiner respectfully disagrees. As discussed above, the claims are not drawn to the specific protease domains of specific MTSP type II, but to polypeptides consisting of any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. The claimed genera of

the claims are drawn to species which are widely variant in structure. The genus of the claims are structurally diverse as it encompasses any catalytically active protease domains of any or all MTSP or MTSP1, excepting having serine protease activity. As such, neither the description of solely structural features present in all members of the genus is sufficient to be representative of the attributes and features of the entire genus.

Hence the rejection is maintained.

Claims 1, 11, 20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 112, first paragraph, because the specification, while being enabling for a polypeptide consisting of amino acids 615-855 of SEQ ID NO:2, does not reasonably provide enablement for a polypeptide consisting of any protease domain of any type II membrane type serine protease (MTSP) or MTSP1 or a catalytically active portion thereof. The specification does not enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the invention commensurate in scope with these claims.

Factors to be considered in determining whether undue experimentation is required are summarized in In re Wands 858 F.2d 731, 8 USPQ2nd 1400 (Fed. Cir. 1988). They include (1) the quantity of experimentation necessary, (2) the amount of direction or guidance presented, (3) the presence or absence of working examples, (4) the nature of the invention, (5) the state of the prior art, (6) the relative skill of those in the art, (7) the predictability or unpredictability of the art, and (8) the breadth of the claims.

Claims 1, 11, 20, 35-36, 40-42 and 113-114 are drawn to a polypeptide consisting of a protease domain or catalytically active fragment thereof of a type-II membrane-type serine protease (MTSP) from any source. Claims 11 and 34 limit the MTSP polypeptide to a MTSP1 polypeptide from any source. Therefore, these claims are drawn to polypeptides having undefined structure.

The scope of the claims is not commensurate with the enablement provided by the disclosure with regard to the extremely large number of polypeptides comprising a protease or catalytically active domain broadly encompassed by the claims. Since the amino acid sequence of a protein determines its structural and functional properties, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a knowledge of and guidance with regard to which amino acids in the protein's sequence, if any, are tolerant of modification and which are conserved (i.e. expectedly intolerant to modification), and detailed knowledge of the ways in which the proteins' structure relates to its function. However, in this case the disclosure is limited to the polypeptide comprising amino acids 615-855 of SEQ ID NO:2, or the amino acids of SEQ ID NO:50.

It would require undue experimentation of the skilled artisan to make and use the claimed polypeptides. The specification is limited to teaching the use of polypeptide consisting of amino acids 615-855 of SEQ ID NO:2 or the amino acids of SEQ ID NO:50 but provides no guidance with regard to the making of variants and mutants or with regard to other uses. In view of the great breadth of the claim, amount of experimentation required to make the claimed polypeptides, the lack of guidance,

working examples, and unpredictability of the art in predicting function from a polypeptide primary structure, the claimed invention would require undue experimentation. As such, the specification fails to teach one of ordinary skill how to use the full scope of the polypeptides encompassed by the claims.

While enzyme isolation techniques, recombinant and mutagenesis techniques are known, and it is routine in the art to screen for multiple substitutions or multiple modifications as encompassed by the instant claims, the specific amino acid positions within a protein's sequence where amino acid modifications can be made with a reasonable expectation of success in obtaining the desired activity/utility are limited in any protein and the result of such modifications is unpredictable. In addition, one skilled in the art would expect any tolerance to modification for a given protein to diminish with each further and additional modification, e.g. multiple substitutions.

The specification does not support the broad scope of the claims which encompass all modifications and variants of a protease or catalytically active domain or modifications of amino acids 615-855 of SEQ ID NO:2 because the specification does not establish: (A) regions of the protein structure which may be modified without affecting MTSP/serine protease activity; (B) the general tolerance of MTSP to modification and extent of such tolerance; (C) a rational and predictable scheme for modifying any amino acid residue with an expectation of obtaining the desired biological function; and (D) the specification provides insufficient guidance as to which of the essentially infinite possible choices is likely to be successful.

Thus, applicants have not provided sufficient guidance to enable one of ordinary skill in the art to make and use the claimed invention in a manner reasonably correlated with the scope of the claims broadly including protease or catalytically active domains of MTSP with an enormous number of amino acid modifications of the MTSP polypeptides and of amino acids 615-855 of SEQ ID NO:2. The scope of the claims must bear a reasonable correlation with the scope of enablement (*In re Fisher*, 166 USPQ 19 24 (CCPA 1970)). Without sufficient guidance, determination of the serine protease domain or the catalytically active domain of MTSP having the desired biological characteristics is unpredictable and the experimentation left to those skilled in the art is unnecessarily, and improperly, extensive and undue. See *In re Wands* 858 F.2d 731, 8 USPQ2nd 1400 (Fed. Cir, 1988).

In response to the previous Office Action, applicants have traversed the above rejection.

Applicants argue that the claims are enabled because the level of skill in the art is high and the specification teaches that MTSP polypeptides constitute a recognized well-known and well characterized family of serine protease and the specification describes the protease domain of a number of MTSP family members, such as conserved features of MTSP protease domains. Examiner respectfully disagrees. The scope of the claims, which are drawn to polypeptides consisting of any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1, is not commensurate with the enablement provided by the

disclosure with regard to the extremely large number of polypeptides comprising a protease or catalytically active domain broadly encompassed by the claims. Even though the structure of some MTSP are known, the claims are drawn to any or all serine domains and catalytically active fragments of any or all protease domains of any or all MTSP or MTSP1. As discussed above, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a specific knowledge of and guidance with regard to which specific amino acids in the protein's sequence, can be modified such that the modified polypeptide continues to have said claimed activity. It is this specific guidance that applicants do not provide. While the art may teach in general the structure of MTSP conserved amino acid sequences, protease domains, X-ray crystal structure and etc, such teachings will not reduce the burden of undue experimentation on those of ordinary skill in the art.

Applicants also argue that the claims are enabled because the knowledge, regarding MTSP proteins, of those skilled in the art is high. The Examiner respectfully disagrees. The claims are drawn to polypeptides consisting of any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. Since the amino acid sequence of the protein determines its structural and functional properties, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a knowledge of and guidance with regard to which amino acids in the protein's sequence, if any, are tolerant of modification and which are conserved (i.e. expectedly intolerant to modification), and

detailed knowledge of the ways in which the proteins' structure relates to its function. In addition, the art does not provide any teaching or guidance as to which amino acids within a serine protease can be modified and which ones are conserved such that one of skill in the art can make the recited polypeptides having serine protease activity and the general tolerance of serine proteases to structural modifications and the extent of such tolerance. The art clearly teaches that changes in a protein's amino acid sequence to obtain the desired activity without any guidance/knowledge as to which amino acids in a protein are required for that activity is highly unpredictable. At the time of the invention, there was a high level of unpredictability associated with altering a polypeptide sequence with an expectation that the polypeptide will maintain the desired activity. For example, Branden et al. (Introduction to Protein Structure, Garland Publishing Inc., New York, page 247, 1991 – cited previously on form PTO-892) teach that (1) protein engineers are frequently surprised by the range of effects caused by single mutations that they hoped would change only one specific and simple property in enzymes, (2) the often surprising results obtained by experiments where single mutations are made reveal how little is known about the rules of protein stability, and (3) the difficulties in designing de novo stable proteins with specific functions.

Applicants argue that the specification discloses working examples, thus a person skilled in the art has sufficient guide in making the claimed polypeptides. Examiner respectfully disagrees. Even though the structure of some MTSP are taught, the claims are not only drawn to polypeptides consisting of catalytically active fragments of only MTSP1, MTSP3, MTSP4 and MTSP6, but to any or all mutants, variants and

recombinants of any MTSP. Without specific guidance, those skilled in the art will be subjected to undue experimentation of making and testing each of the enormously large number of mutants that results from such experimentation. While the art may teach in general the structure of MTSP, conserved amino acid sequences, and etc, such teachings will not reduce the burden of undue experimentation on those of ordinary skill in the art.

Hence the rejection is maintained.

Claim Rejections - 35 USC § 102

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

(b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.

(e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

Claims 1-3 and 19-20 were rejected under 35 U.S.C. **102(b)** as being anticipated by Dawson et al.

In view of the fact that Dawson et al. do not teach an isolated serine protease domain of a MTSP protein, the rejection has been **withdrawn**.

Claims 1, 11-13, 20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. **102(b)** as being anticipated by Takeuchi et al.

Claims 1, 11-13, 20 and 34 are drawn to a polypeptide consisting of a serine protease domain of MTSP having the characteristics recited in the claims. Claims 35-36 are drawn to a conjugate comprising a polypeptide comprising a serine protease domain of MTSP and a targeting agent. Claims 40-42 and 113-114 are drawn to a solid support comprising a polypeptide comprising a serine protease domain of MTSP.

Takeuchi et al. (Reference IJ : PTO-1449) teaches a polypeptide comprising a fragment consisting of a serine protease domain that is 100% identical to amino acids 615-855 of SEQ ID NO:2 of the instant invention (page 11060, 2nd full paragraph). Takeuchi et al. discloses a purified activated protease domain, comprising amino acids 615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VVGGT sequence (Figure 3 and right column on page 11057).

Takeuchi et al. teaches a catalytically active polypeptide comprising the serine protease domain linked to a His-tag (page 11055, 3rd full paragraph, page 11057, 4th full paragraph). Takeuchi et al. also teaches a solid support comprising said polypeptide (page 11057, 4th full paragraph and Figure 5). Therefore, the teaching of Takeuchi et al. anticipates claims 1, 11-13, 20, 34-36, 40-42 and 113-114.

Examiner notes that the contents of the reference were made public at the National Academy of Sciences colloquium held February 20-21, 1999 (see top of reference).

In response to the previous Office Action, applicants have traversed the above rejections.

Applicants argue that Takeuchi et al. does not anticipate the instant claims because the instant claims are drawn to a polypeptide that consists of a protease domain or catalytically active portion thereof. Examiner respectfully disagrees. In addition to the full-length MT-SP1, Takeuchi et al. also discloses a polypeptide consisting of the serine protease domain. The serine protease domain is initially expressed in *E. coli* as a His-tagged fusion, but a renatured active protein lacking the His tag was isolated and N-terminal sequencing of this protein yielded VGGT, which corresponds to residues 615-619 of SEQ ID NO:2 of the instant invention. Takeuchi et al. discloses that Cys at position 731 forms a disulfide bond with Cys 604 present in the pro domain (page 11060). Since the serine protease domain of Takeuchi et al. lacks the pro domain of the wildtype protein, Cys residue at position 731 of said serine protease domain does not form a disulfide bond and therefore is a "free cysteine". The specification on page 58 states that in "the single chain form, the residue at 731 in the protease domain is free" (page 58, lines 15-16). Therefore, the serine protease domain of Takeuchi et al. is a single chain polypeptide.

Applicants also argue that the claims are not anticipated by Takeuchi et al. because Takeuchi et al. does not disclose replacing a free Cys residue of the serine

protease domain of an MTSP polypeptide with another amino acid or a serine residue. Examiner respectfully disagrees. The limitation "a free Cys in the protease domain is replaced with another amino acid" and "a free Cys in the protease domain is replaced with a serine" is a product-by-process type limitation. The end result of the products of the claims is a serine protease domain or a serine protease domain having a serine residue. Whether the product of the claimed protein is obtained by replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not. Therefore, there is no there a structure implied by said limitations. Since the polypeptide of Takeuchi et al. consists of a protease domain of a MTSP and the MTSP protease domain has serine protease activity, the claims are anticipated by the prior art. Also, since the serine protease domain of Takeuchi et al. has a serine residue, claim 20 is also anticipated.

Hence the rejections are maintained.

Claim Rejections - 35 USC § 102/103

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

(e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

The following is a quotation of 35 U.S.C. 103(a), which forms the basis for all obviousness rejections, set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

Claims 1, 11-13 and 34 rejected under 35 U.S.C. 103(a) as obvious over O'Brien et al.

Claims 1, 11-13 and 34 are drawn to a polypeptide comprising a serine protease domain of MTSP.

O'Brien et al. (U.S. Patent No. 5,972,616 – reference P- PTO 1449) teaches a polypeptide having 100% identity to the full length MTSP1 of SEQ ID NO:2 of the instant invention (SEQ ID NO:2, columns 19-24). O'Brien et al. teaches a serine protease domain having proteolytic activity that is 100% identical to amino acids 615-855 of SEQ ID NO:2 (Figure 2, Figure 10 and SEQ ID NO:14). Further, O'Brien et al. teaches a method of expressing polypeptides via a vector in host cells. O'Brien et al. also teaches that the protease domain could be released and be used as a diagnostic which has the potential for a target for therapeutic intervention (Column 15, lines 35-38). Therefore, it would have been obvious to one having ordinary skill in the art at the time the invention was made to express the protease domain of SQ ID NO:14 and purify the polypeptide. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. One of ordinary skill in the art would have had a reasonable expectation of success since expression of a heterologous polypeptide is routine in the art and O'Brien et al. teaches how to express heterologous polypeptides.

Therefore, the above reference renders claims 1, 11-13 and 34 *prima facie* obvious to one of ordinary skill in the art.

In response to the previous Office Action, applicants have traversed the above rejections.

Applicants also argue that one of skill in the art would recognize the disclosure of the polypeptide of O'Brien as not disclosing a single chain polypeptide. Examiner respectfully disagrees. Takeuchi et al. discloses that Cys at position 731 forms a disulfide bond with Cys 604 present in the pro domain (page 11060). Since the serine protease domain of Takeuchi et al. lacks the pro domain of the wildtype protein, Cys residue at position 731 of said serine protease domain does not form a disulfide bond and therefore is a "free cysteine". The specification on page 58 states that in "the single chain form, the residue at 731 in the protease domain is free" (page 58, lines 15-16). Therefore, the serine protease domain of O'Brien et al. is a single chain polypeptide.

Applicants also argue that the claims are not anticipated by O'Brien et al. because O'Brien et al. does not disclose replacing a free Cys residue of the serine protease domain of an MTSP polypeptide with another amino acid. Examiner respectfully disagrees. The limitation "a free Cys in the protease domain is replaced with another amino acid" is a product-by-process type limitation. The end result of the products of the claims is a serine protease domain. Whether the product of the claimed protein is obtained by replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not. Therefore, there is no structure implied by said limitations. Since the

polypeptide of O'Brien et al. consists of a protease domain of a MTSP and the MTSP protease domain has serine protease activity, the claims are anticipated by the prior art.

Applicants also argue that O'Brien et al. provides no teaching or suggestion of smaller fragments having serine protease activity because it does not teach how to make a single chain polypeptide that has serine protease activity. Examiner respectfully disagrees. O'Brien et al. teaches a method of expressing polypeptides via a vector in host cells. It is well within the skill available in the art to purify the protease domain since O'Brien et al. identifies the protease domain. Therefore, it would have been obvious to one having ordinary skill in the art at the time the invention was made to express the protease domain of SQ ID NO:14 and purify the polypeptide. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. One of ordinary skill in the art would have had a reasonable expectation of success since expression of a heterologous polypeptide is routine in the art and O'Brien et al. teaches how to express heterologous polypeptides. Further, since the serine protease domain of Takeuchi et al. lacks the pro domain of the wildtype protein, Cys residue at position 731 of said serine protease domain does not form a disulfide bond and therefore is a "free cysteine". The specification on page 58 states that in "the single chain form, the residue at 731 in the protease domain is free" (page 58, lines 15-16). Also, as discussed previously, the limitation "a free Cys in the protease domain is replaced with another amino acid" is a product-by-process type limitation. The end result of the products of the claims is a serine protease domain. Whether the product of the claimed protein is obtained by

replacing a free cysteine residue or not, the product is still the same because the instant claims may be produced by the recited modification or not. Therefore, there is no there a structure implied by said limitations. Therefore, the serine protease domain of O'Brien et al. is a single chain polypeptide.

Hence the rejections are maintained.

Claims 35-36, 40-42 and 113-114 are rejected under 35 U.S.C. 103(a) as being unpatentable over O'Brien et al.

Claims 35-36 are drawn to a conjugate comprising a polypeptide comprising a serine protease domain of MTSP and a targeting agent. Claims 40-42 and 113-114 are drawn to a solid support comprising a polypeptide comprising a serine protease domain of MTSP.

O'Brien et al. (U.S. Patent No. 5,972,616 – reference P- PTO 1449) teaches a polypeptide having 100% identity to the full length MTSP1 of SEQ ID NO:2 of the instant invention, as discussed above. O'Brien et al. also teaches that the protease domain could be released the used as a diagnostic which has the potential for a target for therapeutic intervention (Column 15, lines 35-38).

O'Brien et al. also teaches method of making fragments of SEQ ID NO:2 (Column 9, lines 22-55). O'Brien et al. teaches said fragments linked to another polypeptide (Column 9, lines 54-55) and conjugated to bridging molecules (Column 6,

lines 27-39) for detecting the polypeptide. Assays using polypeptides linked to the molecules taught by O'Brien et al. utilize solid supports.

Therefore, it would have been obvious to one having ordinary skill in the art at the time the claimed invention was made to make a polypeptide comprising of the serine protease domain of SEQ ID NO:2 taught by O'Brien et al. and to make conjugates and solid support comprising of a polypeptide comprised of the serine protease domain of SEQ ID NO:2. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. The motivation of making conjugates and solid supports comprising of said polypeptide is to use the conjugate and solid support in a variety of diagnostic assays. One of ordinary skill in the art would have had a reasonable expectation of success making fragments of a polypeptide is routine in the art and O'Brien et al. teaches how to make fragments of SEQ ID NO:2. One of ordinary skill in the art would have had a reasonable expectation of success in diagnostic assays using conjugates and solid supports comprising a polypeptide is very well known, as taught by O'Brien et al.

Therefore, the above references render claims 35-36 and 40-42 *prima facie* obvious to one of ordinary skill in the art.

In response to the previous Office Action, applicants have traversed the above rejection and has been discussed above.

Hence the rejection is maintained.

The rejection of claims 19-20 under 35 U.S.C. 103(a) as being unpatentable over O'Brien et al. and Estell et al. in view of Takeuchi et al. has been withdrawn.

Conclusion

None of the claims are in condition for allowance.

THIS ACTION IS MADE FINAL. Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire **THREE MONTHS** from the mailing date of this action. In the event a first reply is filed within **TWO MONTHS** of the mailing date of this final action and the advisory action is not mailed until after the end of the **THREE-MONTH** shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than **SIX MONTHS** from the mailing date of this final action.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Yong Pak whose telephone number is 571-272-0935. The examiner can normally be reached 6:30 A.M. to 5:00 P.M. Monday through Thursday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Nashaat Nashed can be reached on 571-272-0934. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Any inquiry of a general nature or relating to the status of this application or proceeding should be directed to the receptionist whose telephone number is 571-272-1600.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should

Application/Control Number: 09/776,191

Page 25

Art Unit: 1652

you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll free).

/Yong D Pak/

Primary Examiner, Art Unit 1652

Exhibit 2



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/776,191	02/02/2001	Edwin L. Madison	17106-017001 / 1607	3237
20985 7590 06/25/2007 FISH & RICHARDSON, PC P.O. BOX 1022 MINNEAPOLIS, MN 55440-1022			EXAMINER PAK, YONG D	
			ART UNIT 1652	PAPER NUMBER
			MAIL DATE 06/25/2007	DELIVERY MODE PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary

Application No.

09/776,191

Applicant(s)

MADISON ET AL.

Examiner

Yong D. Pak

Art Unit

1652

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 23 March 2007.
- 2a) ☐ This action is FINAL. 2b) ☒ This action is non-final.
- 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) See Continuation Sheet is/are pending in the application.

4a) Of the above claim(s) 10, 43-46, 48-55, 108, 109, 115, 116, 118-120 and 122-126 is/are withdrawn from consideration.

- 5) ☐ Claim(s) _____ is/are allowed.
- 6) ☒ Claim(s) 1-3, 11-13, 19, 20, 34-36, 40-42, 113 and 114 is/are rejected.
- 7) ☐ Claim(s) _____ is/are objected to.
- 8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☐ The specification is objected to by the Examiner.
- 10) ☐ The drawing(s) filed on _____ is/are: a) ☐ accepted or b) ☐ objected to by the Examiner.
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
a) ☐ All b) ☐ Some * c) ☐ None of:
1. ☐ Certified copies of the priority documents have been received.
 2. ☐ Certified copies of the priority documents have been received in Application No. _____.
 3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- | | |
|--|---|
| 1) <input checked="" type="checkbox"/> Notice of References Cited (PTO-892) | 4) <input type="checkbox"/> Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____ |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948) | 5) <input type="checkbox"/> Notice of Informal Patent Application |
| 3) <input type="checkbox"/> Information Disclosure Statement(s) (PTO/SB/08)
Paper No(s)/Mail Date _____ | 6) <input type="checkbox"/> Other: _____ |

Continuation of Disposition of Claims: Claims pending in the application are 1-3, 10-13, 19, 20, 34-36, 40-46, 48-55, 108, 109, 113-116, 118-120 and 122-126.

Application/Control Number:
09/776,191
Art Unit: 1652

Page 2

DETAILED ACTION

The petition of March 23, 2007 is being treated as a request for reconsideration. In view of said request, the finality of the previous Office action is withdrawn, rendering the petition moot. A new action on the merits is set forth below.

This application is a CIP of 09/657,986, now issued as U.S. Patent No. 6,797,504.

The amendment filed on October 23, 2006, amending claims 1, 12, 13 and 19 and canceling claim 5, has been entered.

Claims 1-3, 10-13, 19-20, 34-36, 40-46, 48-55, 108-109 113-116, 118-120 and 122-126 are pending. Claims 10, 43-46, 48-55, 108-109, 115-116, 118-120 and 122-126 are withdrawn. Claims 1-3, 11-13, 19-20, 34-36, 40-42 and 113-114 are under consideration.

Priority

Applicant's claim for domestic priority under 35 U.S.C. 119(e) is acknowledged. However, the provisional applications upon which priority is claimed fails to provide adequate support under 35 U.S.C. 112 for claims 11-13 and 34 of this application.

Provisional applications 60/179,982, 60/183,542, 60/213,124, 60/220,970 and 60/234,840 fail to provide adequate support for polypeptides comprising the serine protease domain of MTSP1. Provisional applications 60/179,982 and 60/183,542

Application/Control Number:
09/776,191
Art Unit: 1652

Page 3

describe polypeptides related MTSP3 and provisional application 60/213,124,
60/220,970 and 60/234,840 describe polypeptides related to MTSP4.

Therefore, the effective filing date for purpose of prior art is the filing date of
09/657,986, which is 9/8/2000.

Response to Arguments

Applicant's amendment and arguments filed on October 23, 2006, have been
fully considered and are deemed to be persuasive to overcome the rejections previously
applied. Rejections and/or objections not reiterated from previous office actions are
hereby withdrawn.

Claim Objections

Claims 11-13 and 34 are objected for being drawn to non-elected subject matter.
In response to the previous Office Action, applicants have traversed the above rejection.
Applicants argue that claims 11-13 and 34 should be retained pending a determination
of the allowability of claim 1, which is a linking claim, linking the elected subject matter.
Since claim 1 has not been indicated as allowable, the objection is maintained.

Claim Rejections - 35 USC § 112

The following is a quotation of the second paragraph of 35 U.S.C. 112:

The specification shall conclude with one or more claims particularly pointing out and distinctly
claiming the subject matter which the applicant regards as his invention.

Claims 1-3, 11-12, 13 and claims 19-20, 34-36, 40-42 and 113-114 depending therefrom rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claims 1-3, 11-12, 13 recite the phrase "substantially purified single-chain polypeptide". The metes and bounds of the phrase in the context of the above claims are not clear to the Examiner. It is not clear to the Examiner what is considered as "substantially purified" by the applicants. A perusal of the specification did not provide a clear definition for the above phrase. Without a clear definition, those skilled in the art would be unable to conclude if a polypeptide is a "substantially purified" polypeptide without knowing the metes and bounds of the phrase. Examiner requests clarification of the above phrase.

In response to the previous Office Action, applicants have traversed the above rejection.

Applicants argue that when read in light of the specification, the skilled artisan would understand the meaning of the recitation "substantially purified" and points to page 46, lines 4-15 of the specification for the definition of the phrase "substantially purified". Examiner respectfully disagrees. The specification on page 46, lines 4-15, does not define what applicants mean by "substantially purified", but only describes that "substantially pure means sufficiently homogeneous to appear free of readily detectable impurities as determined by standard methods of analysis". Since there is no clear guidance to one having ordinary skill in the art in qualifying the purity of an enzyme by

ascertaining whether it is free of readily detectable impurities, it is not clear to the Examiner as to how much of a presence of these readily detectable impurities qualifies an enzyme to be "substantially pure". Therefore, those skilled in the art would be unable to conclude what polypeptides are "substantially purified".

Hence the rejection is maintained.

The following is a quotation of the first paragraph of 35 U.S.C. 112:

The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

Claims 1-3, 11, 19-20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

Claims 1-3, 11, 19-20, 35-36, 40-42 and 113-114 are drawn to a polypeptide consisting of a protease domain or catalytically active fragment thereof of type-II membrane-type serine protease (MTSP) from any source. Claims 11 and 34 limit the MTSP polypeptide to a MTSP1 polypeptide from any source. Therefore, these claims are drawn to a genus of polypeptides having any structure. The specification only teaches four species, amino acids 615-855 of SEQ ID NO:2 (MTSP1), amino acids of 205-437 of SEQ ID NO:4 (MTSP3), amino acids of SEQ ID NO:6 (MTSP4) and amino acids 217-443 of SEQ ID NO:11 (MTSP6). These species are not enough to describe

the whole genus and there is no evidence on the record of the relationship between the structure of the above catalytically active protease domains of SEQ ID NOs: 2, 4, 6 and 11 and the structure of the serine protease domain of any or all MTSP polypeptides or MTSP1 polypeptides. Further, the specification does not describe the structure of a catalytically active fragment of a protease domain of any or all MTSP polypeptide. Therefore, the specification fails to describe a representative species of the genus of polypeptides comprising of a serine protease domain or a catalytically active portion of a MTSP polypeptide.

Given this lack of description of the representative species encompassed by the genus of the claims, the specification fails to sufficiently describe the claimed invention in such full, clear, concise, and exact terms that a skilled artisan would recognize that applicants were in possession of the inventions of claims 1-3, 11, 19-20, 34-36, 40-42 and 113-114.

Applicant is referred to the revised guidelines concerning compliance with the written description requirement of U.S.C. 112, first paragraph, published in the Official Gazette and also available at www.uspto.gov.

In response to the previous Office Action, applicants have traversed the above rejection.

Applicants argue that the claims are fully described by the specification because the structural feature, a single chain protease domain, is present in all members of the genus and is the defining and requisite property and the specification clearly describes

this feature. Examiner respectfully disagrees. The recitation of "protease domain of a MTSP" or "MTSP1" fails to provide a sufficient description of the claimed genus of polynucleotides as it merely describes the functional features of the genus without providing any definition of the structural features of the species within the genus. The CAFC in *UC California v. Eli Lilly*, (43 USPQ2d 1398) stated that: "in claims to genetic material, however a generic statement such as 'vertebrate insulin cDNA' or 'mammalian insulin cDNA,' without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function. It does not specifically define any of the genes that fall within its definition. It does not define any structural features commonly possessed by members of the genus that distinguish them from others. One skilled in the art therefore cannot, as one can do with a fully described genus, visualize or recognize the identity of the members of the genus." Similarly with the claimed genus of protease domains, the functional definition of the genus does not provide any structural information commonly possessed by members of the genus which distinguish the species within the genus from other proteins such that one can visualize or recognize the identity of the members of the genus.

Applicants also argue that the claims are fully described because the specification describes known MTSPs and identifies the protease domains thereof, unknown MTSPs and its protease domains. Examiner respectfully disagrees. The claims are not limited to specific protease domains of specific MTSP proteins, but the claims are drawn to polypeptides comprising any protease domains or any or all

catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. As discussed in the written description guidelines, the written description requirement for a claimed genus may be satisfied through sufficient description of a representative number of species by actual reduction to practice, reduction to drawings, or by disclosure of relevant, identifying characteristics, i.e., structure or other physical and/or chemical properties, by functional characteristics coupled with a known or disclosed correlation between function and structure, or by a combination of such identifying characteristics, sufficient to show the applicant was in possession of the claimed genus. A representative number of species means that the species which are adequately described are representative of the entire genus. **Thus, when there is substantial variation within the genus, one must describe a sufficient variety of species to reflect the variation within the genus.** Satisfactory disclosure of a representative number depends on whether one of skill in the art would recognize that the applicant was in possession of the necessary common attributes or features of the elements possessed by the members of the genus in view of the species disclosed. For inventions in an unpredictable art, adequate written description of a genus which embraces widely variant species cannot be achieved by disclosing only one species within the genus. In the instant case the claimed genera of the claims are drawn to species which are widely variant in structure. The genus of the claims are structurally diverse as it encompasses any catalytically active protease domains of any or all MTSP or MTSP1, excepting having serine protease activity. As such, neither the description of

solely structural features present in all members of the genus is sufficient to be representative of the attributes and features of the entire genus.

Applicants also argue that the specification provides "relevant, identifying characteristics" of a representative number of species of the claimed genus. Examiner respectfully disagrees. The claims are drawn to polypeptides comprising any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. The claims are drawn to polypeptides having any structure and therefore, the claims are drawn to a genus encompassing species having substantial variation and fails to describe a representative number of species. As discussed in the written description guidelines, the written description requirement for a claimed genus may be satisfied through sufficient description of a representative number of species by actual reduction to practice, reduction to drawings, or by disclosure of relevant, identifying characteristics, i.e., structure or other physical and/or chemical properties, by functional characteristics coupled with a known or disclosed correlation between function and structure, or by a combination of such identifying characteristics, sufficient to show the applicant was in possession of the claimed genus. A representative number of species means that the species which are adequately described are representative of the entire genus. **Thus, when there is substantial variation within the genus, one must describe a sufficient variety of species to reflect the variation within the genus.** Satisfactory disclosure of a representative number depends on whether one of skill in the art would recognize that the applicant

was in possession of the necessary common attributes or features of the elements possessed by the members of the genus in view of the species disclosed. For inventions in an unpredictable art, adequate written description of a genus which embraces widely variant species cannot be achieved by disclosing only one species within the genus. In the instant case the claimed genera of the claims are drawn to species which are widely variant in structure. The genus of the claims are structurally diverse as it encompasses any catalytically active protease domains of any or all MTSP or MTSP1, excepting having serine protease activity. As such, neither the description of solely structural features present in all members of the genus is sufficient to be representative of the attributes and features of the entire genus.

Applicants also argue that the claims are fully described because specification provides at least a dozen examples of protease domains of MTSPs. Examiner respectfully disagrees. The claims are not drawn to the specific protease domains of the MTSPs disclosed in the specification, but to polypeptides consisting of any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. In view of the widely variant species encompassed by the genus, the species disclosed in the specification is not enough and does not constitute a representative number of species to describe the whole genus of any or all variants, recombinant and mutants of any or all polypeptides having serine protease activity isolated from any or all source, including any or all variants, recombinants and mutants thereof, and there is no evidence on the record of the relationship between the structure

of the protease domain of the specific MTSPs disclosed in the specification and the structure of any or all recombinant, variant and mutant of any or all polypeptides having serine protease activity. Therefore, the specification fails to describe a representative species of the genus comprising any or all polypeptides having serine protease activity, including any or all variants, recombinants and mutants thereof.

Hence the rejection is maintained.

Claims 1-3, 11, 19-20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 112, first paragraph, because the specification, while being enabling for a polypeptide consisting of amino acids 615-855 of SEQ ID NO:2, does not reasonably provide enablement for a polypeptide comprising any protease domain of any type II membrane type serine protease (MTSP) or MTSP1 or a catalytically active portion thereof. The specification does not enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the invention commensurate in scope with these claims.

Factors to be considered in determining whether undue experimentation is required are summarized in In re Wands 858 F.2d 731, 8 USPQ2nd 1400 (Fed. Cir. 1988). They include (1) the quantity of experimentation necessary, (2) the amount of direction or guidance presented, (3) the presence or absence of working examples, (4) the nature of the invention, (5) the state of the prior art, (6) the relative skill of those in the art, (7) the predictability or unpredictability of the art, and (8) the breadth of the claims.

Claims 1-3, 11, 19-20, 35-36, 40-42 and 113-114 are drawn to a polypeptide consisting of a protease domain or catalytically active fragment thereof of a type-II membrane-type serine protease (MTSP) from any source. Claims 11 and 34 limit the MTSP polypeptide to a MTSP1 polypeptide from any source. Therefore, these claims are drawn to polypeptides having undefined structure.

The scope of the claims is not commensurate with the enablement provided by the disclosure with regard to the extremely large number of polypeptides comprising a protease or catalytically active domain broadly encompassed by the claims. Since the amino acid sequence of a protein determines its structural and functional properties, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a knowledge of and guidance with regard to which amino acids in the protein's sequence, if any, are tolerant of modification and which are conserved (i.e. expectedly intolerant to modification), and detailed knowledge of the ways in which the proteins' structure relates to its function. However, in this case the disclosure is limited to the polypeptide comprising amino acids 615-855 of SEQ ID NO:2, or the amino acids of SEQ ID NO:50.

It would require undue experimentation of the skilled artisan to make and use the claimed polypeptides. The specification is limited to teaching the use of polypeptide comprising amino acids 615-855 of SEQ ID NO:2 or the amino acids of SEQ ID NO:50 but provides no guidance with regard to the making of variants and mutants or with regard to other uses. In view of the great breadth of the claim, amount of experimentation required to make the claimed polypeptides, the lack of guidance,

working examples, and unpredictability of the art in predicting function from a polypeptide primary structure, the claimed invention would require undue experimentation. As such, the specification fails to teach one of ordinary skill how to use the full scope of the polypeptides encompassed by the claims.

While enzyme isolation techniques, recombinant and mutagenesis techniques are known, and it is routine in the art to screen for multiple substitutions or multiple modifications as encompassed by the instant claims, the specific amino acid positions within a protein's sequence where amino acid modifications can be made with a reasonable expectation of success in obtaining the desired activity/utility are limited in any protein and the result of such modifications is unpredictable. In addition, one skilled in the art would expect any tolerance to modification for a given protein to diminish with each further and additional modification, e.g. multiple substitutions.

The specification does not support the broad scope of the claims which encompass all modifications and variants of a protease or catalytically active domain or modifications of amino acids 615-855 of SEQ ID NO:2 because the specification does not establish: (A) regions of the protein structure which may be modified without affecting MTSP/serine protease activity; (B) the general tolerance of MTSP to modification and extent of such tolerance; (C) a rational and predictable scheme for modifying any amino acid residue with an expectation of obtaining the desired biological function; and (D) the specification provides insufficient guidance as to which of the essentially infinite possible choices is likely to be successful.

Thus, applicants have not provided sufficient guidance to enable one of ordinary skill in the art to make and use the claimed invention in a manner reasonably correlated with the scope of the claims broadly including protease or catalytically active domains of MTSP with an enormous number of amino acid modifications of the MTSP polypeptides and of amino acids 615-855 of SEQ ID NO:2. The scope of the claims must bear a reasonable correlation with the scope of enablement (*In re Fisher*, 166 USPQ 19 24 (CCPA 1970)). Without sufficient guidance, determination of the serine protease domain or the catalytically active domain of MTSP having the desired biological characteristics is unpredictable and the experimentation left to those skilled in the art is unnecessarily, and improperly, extensive and undue. See *In re Wands* 858 F.2d 731, 8 USPQ2nd 1400 (Fed. Cir, 1988).

In response to the previous Office Action, applicants have traversed the above rejection.

Applicants argue that the claims are enabled because the level of skill in the art is high and the specification teaches that MTSP polypeptides constitute a recognized well-known and well characterized family of serine protease and the specification describes the protease domain of a number of MTSP family members, such as conserved features of MTSP protease domains. Examiner respectfully disagrees. The scope of the claims, which are drawn to polypeptides comprising any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1, is not commensurate with the enablement provided by the disclosure

with regard to the extremely large number of polypeptides comprising a protease or catalytically active domain broadly encompassed by the claims. Even though the structure of some MTSP are known, the claims are drawn to any or all serine domains and catalytically active fragments of any or all protease domains of any or all MTSP or MTSP1. As discussed above, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a specific knowledge of and guidance with regard to which specific amino acids in the protein's sequence, can be modified such that the modified polypeptide continues to have said claimed activity. It is this specific guidance that applicants do not provide. While the art may teach in general the structure of MTSP conserved amino acid sequences, protease domains, X-ray crystal structure and etc, such teachings will not reduce the burden of undue experimentation on those of ordinary skill in the art.

Applicants also argue that the claims are enabled because the knowledge, regarding MTSP proteins, of those skilled in the art is high. The Examiner respectfully disagrees. The claims are drawn to polypeptides comprising any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. Since the amino acid sequence of the protein determines its structural and functional properties, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a knowledge of and guidance with regard to which amino acids in the protein's sequence, if any, are tolerant of modification and which are conserved (i.e. expectedly intolerant to modification), and

detailed knowledge of the ways in which the proteins' structure relates to its function. In addition, the art does not provide any teaching or guidance as to which amino acids within a serine protease can be modified and which ones are conserved such that one of skill in the art can make the recited polypeptides having serine protease activity and the general tolerance of serine proteases to structural modifications and the extent of such tolerance. The art clearly teaches that changes in a protein's amino acid sequence to obtain the desired activity without any guidance/knowledge as to which amino acids in a protein are required for that activity is highly unpredictable. At the time of the invention, there was a high level of unpredictability associated with altering a polypeptide sequence with an expectation that the polypeptide will maintain the desired activity. For example, Branden et al. (Introduction to Protein Structure, Garland Publishing Inc., New York, page 247, 1991) teach that (1) protein engineers are frequently surprised by the range of effects caused by single mutations that they hoped would change only one specific and simple property in enzymes, (2) the often surprising results obtained by experiments where single mutations are made reveal how little is known about the rules of protein stability, and (3) the difficulties in designing de novo stable proteins with specific functions.

Applicants argue that the specification discloses working examples, thus a person skilled in the art has sufficient guide in making the claimed polypeptides. Examiner respectfully disagrees. Even though the structure of some MTSP are taught, the claims are not only drawn to polypeptides comprising catalytically active fragments of only MTSP1, MTSP3, MTSP4 and MTSP6, but to any or all mutants, variants and

recombinants of any MTSP. Without specific guidance, those skilled in the art will be subjected to undue experimentation of making and testing each of the enormously large number of mutants that results from such experimentation. While the art may teach in general the structure of MTSP, conserved amino acid sequences, and etc, such teachings will not reduce the burden of undue experimentation on those of ordinary skill in the art.

Hence the rejection is maintained.

Claim Rejections - 35 USC § 102

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

(a) the invention was known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for a patent.

(b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.

(e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

Claims 1-3 and 19-20 are rejected under 35 U.S.C. 102(b) as being anticipated by Dawson et al.

Claims 1-3 and 19-20 are drawn to a polypeptide consisting of a serine protease domain of MTSP or catalytically active fragments thereof.

Dawson et al. (US Patent 5,465,833 -form PTO-892) discloses a polypeptide consisting of serine protease domain or a catalytically active fragment thereof of a MTSP protein, hepsin (Figure 1). Therefore, the reference of Dawson et al. anticipates claims 1-3 and 19-20.

Claims 1-3, 11-13, 19-20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 102(b) as being anticipated by Takeuchi et al.

Claims 1-3, 11-13, 19-20 and 34 are drawn to a polypeptide comprising fragment consisting of a serine protease domain of MTSP having the characteristics recited in the claims. Claims 35-36 are drawn to a conjugate comprising a polypeptide comprising a serine protease domain of MTSP and a targeting agent. Claims 40 -42 and 113-114 are drawn to a solid support comprising a polypeptide comprising a serine protease domain of MTSP.

Takeuchi et al. (Reference IJ : PTO-1449) teaches a polypeptide comprising a fragment consisting of a serine protease domain that is 100% identical to amino acids 615-855 of SEQ ID NO:2 of the instant invention (page 11060, 2nd full paragraph). Takeuchi et al. discloses a purified activated protease domain, comprising amino acids 615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VVGGT sequence (Figure 3 and right column on page 11057). The MTSP of Takeuchi et al. is not expressed on normal endothelia cells (page 11054, last paragraph and page 11055, 2nd full paragraph), is of

human origin (Figure 1), consists essentially of the protease domain having catalytic activity (page 11060, 2nd full paragraph), and is expressed in tumor cells (page 11055, top paragraph).

Takeuchi et al. teaches a catalytically active polypeptide comprising the serine protease domain linked to a His-tag (page 11055, 3rd full paragraph, page 11057, 4th full paragraph). Takeuchi et al. also teaches a solid support comprising said polypeptide (page 11057, 4th full paragraph and Figure 5). Therefore, the teaching of Takeuchi et al. anticipates claims 1-3, 11-13, 19-20, 34-36, 40-42 and 113-114.

Examiner notes that the contents of the reference were made public at the National Academy of Sciences colloquium held February 20-21, 1999 (see top of reference).

In response to the previous Office Action, applicants have traversed the above rejections.

Applicants argue that Takeuchi et al. does not anticipate the instant claims because the instant claims are drawn to a polypeptide that consists of a protease domain or catalytically active portion thereof. Examiner respectfully disagrees. In addition to the full-length MT-SP1, Takeuchi et al. also discloses a purified activated protease domain, consisting of amino acids 615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VVGGT sequence (Figure 3 and right column on page 11057). Therefore, said purified, activated protease domain anticipates the instant claims.

Applicants also argue that Takeuchi et al. does not anticipate the instant claims because the claimed polypeptide is a single chain polypeptide. Examiner respectfully disagrees. As discussed above, Takeuchi et al. discloses a purified activated protease domain, consisting of amino acids 615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VGGT sequence (Figure 3 and right column on page 11057).

Hence the rejections are maintained.

Claim Rejections - 35 USC § 102/103

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

(e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

The following is a quotation of 35 U.S.C. 103(a), which forms the basis for all obviousness rejections, set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

Claims 1-3, 11-13 and 34 rejected under 35 U.S.C. 103(a) as obvious over

O'Brien et al.

Claims 1-3, 11-13 and 34 are drawn to a polypeptide comprising a serine protease domain of MTSP.

O'Brien et al. (U.S. Patent No. 5,972,616 – reference P- PTO 1449) teaches a polypeptide having 100% identity to the full length MTSP1 of SEQ ID NO:2 of the instant invention (SEQ ID NO:2, columns 19-24). O'Brien et al. teaches a serine protease domain having proteolytic activity that is 100% identical to amino acids 615-855 of SEQ ID NO:2 (Figure 2, Figure 10 and SEQ ID NO:14). Further, O'Brien et al. teaches a method of expressing polypeptides via a vector in host cells. O'Brien et al. also teaches that the protease domain could be released and used as a diagnostic which has the potential for a target for therapeutic intervention (Column 15, lines 35-38). Therefore, it would have been obvious to one having ordinary skill in the art at the time the invention was made to express the protease domain of SQ ID NO:14 and purify the polypeptide. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. One of ordinary skill in the art would have had a reasonable expectation of success since expression of a heterologous polypeptide is routine in the art and O'Brien et al. teaches how to express heterologous polypeptides.

Therefore, the above reference renders claims 1-3, 11-13 and 34 prima facie obvious to one of ordinary skill in the art.

In response to the previous Office Action, applicants have traversed the above rejections.

Applicants also argue that one of skill in the art would recognize the disclosure of the polypeptide of O'Brien as not disclosing a single chain polypeptide. Examiner respectfully disagrees. A single chain polypeptide is one sequence of amino acids beginning with a carboxyl end and terminating with an amino end, wherein the amino acids are connected via peptide bonds. Therefore, the protease domain obtained from O'Brien et al. can be construed as a single chain polypeptide.

Applicants also argue that O'Brien et al. provides no teaching or suggestion of smaller fragments having serine protease activity because it does not teach how to make a single chain polypeptide that has serine protease activity. Examiner respectfully disagrees. O'Brien et al. teaches a method of expressing polypeptides via a vector in host cells. It is well within the skill available in the art to purify the protease domain since O'Brien et al. identifies the protease domain. Therefore, it would have been obvious to one having ordinary skill in the art at the time the invention was made to express the protease domain of SQ ID NO:14 and purify the polypeptide. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. One of ordinary skill in the art would have had a reasonable expectation of success since expression of a heterologous polypeptide is routine in the art and O'Brien et al. teaches how to express heterologous polypeptides.

Applicants again argue that at the time of filing the instant application, one of skill in the art would not have had a reasonable expectation of success to express the protease domain because art evidences that a single-chained polypeptide would not

have been expected to have protease activity. Examiner respectfully disagrees. The claims are drawn to a polypeptide comprising a fragment consisting of a protease domain of SEQ ID NO:2. Therefore, said polypeptide being a single-chained polypeptide is an inherent property of said polypeptide since two polypeptides having identical structure will have identical function and physical and chemical properties.

Hence the rejections are maintained.

Claims 35-36, 40-42 and 113-114 are rejected under 35 U.S.C. 103(a) as being unpatentable over O'Brien et al.

Claims 35-36 are drawn to a conjugate comprising a polypeptide comprising a serine protease domain of MTSP and a targeting agent. Claims 40-42 and 113-114 are drawn to a solid support comprising a polypeptide comprising a serine protease domain of MTSP.

O'Brien et al. (U.S. Patent No. 5,972,616 – reference P- PTO 1449) teaches a polypeptide having 100% identity to the full length MTSP1 of SEQ ID NO:2 of the instant invention, as discussed above. O'Brien et al. also teaches that the protease domain could be released and used as a diagnostic which has the potential for a target for therapeutic intervention (Column 15, lines 35-38).

O'Brien et al. also teaches method of making fragments of SEQ ID NO:2 (Column 9, lines 22-55). O'Brien et al. teaches said fragments linked to another polypeptide (Column 9, lines 54-55) and conjugated to bridging molecules (Column 6,

lines 27-39) for detecting the polypeptide. Assays using polypeptides linked to the molecules taught by O'Brien et al. utilize solid supports.

Therefore, it would have been obvious to one having ordinary skill in the art at the time the claimed invention was made to make a polypeptide comprising of the serine protease domain of SEQ ID NO:2 taught by O'Brien et al. and to make conjugates and solid support comprising of a polypeptide comprised of the serine protease domain of SEQ ID NO:2. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. The motivation of making conjugates and solid supports comprising of said polypeptide is to use the conjugate and solid support in a variety of diagnostic assays. One of ordinary skill in the art would have had a reasonable expectation of success making fragments of a polypeptide is routine in the art and O'Brien et al. teaches how to make fragments of SEQ ID NO:2. One of ordinary skill in the art would have had a reasonable expectation of success in diagnostic assays using conjugates and solid supports comprising a polypeptide is very well known, as taught by O'Brien et al.

Therefore, the above references render claims 35-36 and 40-42 *prima facie* obvious to one of ordinary skill in the art.

In response to the previous Office Action, applicants have traversed the above rejections. Applicants argue that the teachings of O'Brien et al. does not result in the instantly claimed compositions because O'Brien et al. does not teach or suggest a single chain polypeptide that includes a MTSP protease domain where the polypeptide

does not include any additional MTSP portions and the polypeptide has serine protease activity. O'Brien et al. does teach or suggest a single chain polypeptide comprising a MTSP portion, wherein the MTSP portion is a protease domain and wherein the MTSP portion has serine protease activity and wherein the MTSP portion is the only portion of the polypeptide because O'Brien et al. identifies the serine protease domain and one having ordinary skill in the art at the time the invention was filed would have been motivated to purify the serine protease domain of O'Brien et al. as discussed above.

Hence the rejection is maintained.

Claims 19-20 are rejected under 35 U.S.C. 103(a) as being unpatentable over O'Brien et al. and Estell et al. in view of Takeuchi et al.

Claims 19-20 are drawn to a polypeptide comprising the serine protease domain of a MTSP wherein free Cys residues are substituted with Ser residues.

O'Brien et al. teaches a serine protease domain of a MTSP polypeptide, as discussed above.

The reference of O'Brien et al. does not teach a serine protease domain of a MTPSP polypeptides wherein free Cys residues have been replaced with Ser residues.

It is well known in the art that proteins form disulfide bonds via the SH groups of Cys residues. Upon making a polypeptide comprising a serine protease domain, a Cys residue which normally forms disulfide bonds in the full length polypeptide may be left free. For example, Takeuchi et al. (Reference IJ : PTO-1449) teaches that Cysteine at

position 731 of SEQ ID NO:2 normally forms a disulfide bond with a Cys residue in the pro-protease domain (see page 11060, top left paragraph and Figures 1 and 2).

Cys residues are sensitive to oxidation due to their SH side group. Estell et al. (U.S. Patent No. 5,346,823) teaches that Cys residues replaced with Ser residues to decrease a polypeptide's susceptibility to oxidation (Abstract and Column 10, lines 34-38). Ser residues have similar side chains as Cys residues and substitution of a Cys residue with a Ser residue is a conservative substitution.

Therefore, it would have been obvious to one having ordinary skill in the art at the time the claimed invention was made to replace free Cys residues in the protease domain taught by O'Brien et al. with a Ser residue. One of ordinary skill in the art would be motivated to make such a change in order to enhance stability of the polypeptide. One of ordinary skill in the art would have had a reasonable expectation of success since Estell et al. teaches successful decrease of a protein's susceptibility to oxidation by substituting residues sensitive to oxidation with conservative substitutions.

Therefore, the above references render claims 1 and 16, 18-20, 34 and 137 *prima facie* obvious to one of ordinary skill in the art.

In response to the previous Office Action, applicants have traversed the above rejections. Applicants argue that the combination of the teachings of O'Brien et al. with the teachings of Estell et al., and Takeuchi et al. does not result in the instantly claimed methods because O'Brien et al. does not teach or suggest a single chain polypeptide that includes a MTSP protease domain where the polypeptide does not include any

additional MTSP portions and the polypeptide has serine protease activity and that neither Takeuchi et al. nor Estell et al. remedy the defects of O'Brien et al. First, the claims are product claims and not method claims. Second, O'Brien et al. does teach or suggest a single chain polypeptide comprising a MTSP portion, wherein the MTSP portion is a protease domain and wherein the MTSP portion has serine protease activity and wherein the MTSP portion is the only portion of the polypeptide because O'Brien et al. identifies the serine protease domain and one having ordinary skill in the art at the time the invention was filed would have been motivated to purify the serine protease domain of O'Brien et al. as discussed above.

Applicants argue that Takeuchi et al. teaches that every cysteine residue of the protein is disulfide bonded and therefore Takeuchi et al. does not teach or suggest an MTSP protease domain having a free Cys residue. Examiner respectfully disagrees. Figure 4 applicants are referring to illustrate disulfide bonds of cysteine residues of the full length MTSP, for example, the Cys at position 830 is disulfide bonded to Cys at position 191.

Hence the rejection is maintained.

None of the claims are in condition for allowance.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Yong Pak whose telephone number is 571-272-0935. The examiner can normally be reached 6:30 A.M. to 5:00 P.M. Monday through Thursday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Ponnathapu Achutamurthy can be reached on 571-272-0928. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Application/Control Number:
09/776,191
Art Unit: 1652

Page 28

Any inquiry of a general nature or relating to the status of this application or proceeding should be directed to the receptionist whose telephone number is 571-272-1600.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll free).



Yong D. Pak
Patent Examiner 1652

Notice of References Cited

Application/Control No.

09/776,191

Applicant(s)/Patent Under
Reexamination
MADISON ET AL.

Examiner

Yong D. Pak

Art Unit

1652

Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-5,645,833	07-1997	Dawson et al.	424/94.64
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Branden et al. Introduction to Protein Structure, Garland Publishing Inc., New York, page 247, 1991
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Introduction to Protein Structure

Carl Branden & John Tooze

THE COVER

Front: The background photograph of the cover is of a Laue x-ray diffraction pattern produced by a crystal of the plant enzyme ribulose biphosphate carboxylase. This technique is described in Chapter 17. Information derived from such x-ray patterns, together with a knowledge of the amino acid sequence, enabled the three-dimensional arrangement of atoms in the protein to be determined. A simplified representation of this protein structure is shown in color, superimposed on the diffraction pattern. The enzyme, which is involved in the fixation of carbon dioxide, is a member of the large class of α/β barrel protein structures. This class of structures is discussed in detail in Chapter 4.

Back: Tomato bushy stunt virus is a spherical virus made from 180 protein subunits. Arms extending from sixty of these subunits contribute to an internal framework that determines the size of the correctly assembled virus particle. The interdigitated arms from three subunits meet at each of the twenty icosahedral threefold axes of the virus. One such axis is shown here with the β strands from three subunits shown in different shades of green. Virus structure is described in more detail in Chapter 11.

© 1991 Carl Branden and John Tooze

All rights reserved. No part of this book covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems—without permission of the publisher.

Library of Congress Cataloging-in-Publication Data

Branden, Carl.

Introduction to protein structure / Carl Branden, John Tooze.

p. cm.

Includes index.

ISBN 0-8153-0344-0 — ISBN 0-8153-0270-3 (pbk.)

1. Proteins—Structure. I. Tooze, John. II. Title.

QP551.B7635 1991

574.19'245—dc20

91-11788

CIP

Published by Garland Publishing, Inc.

136 Madison Ave., New York, New York, 10016

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Prediction, Engineering, and Design of Protein Structures

16

Over a period of more than 3 billion years a large variety of protein molecules has evolved to run the complex machinery of present-day cells and organisms. Most of us believe that these molecules have evolved by random mutation of genes and natural selection for those gene products that have conferred some functional advantage contributing to the survival of individual organisms.

Long before Darwin and Wallace proposed the theory of evolution and Mendel discovered the laws of genetics, plant and animal breeders had begun to interfere with the process of evolution in the species that gave rise to domesticated animals and cultivated plants. Considering their total lack of knowledge of both evolutionary theory and genetics, their achievements, brought about by forcing the pace of and subverting natural selection, were impressive albeit very gradual. With the advent of molecular genetics and in particular techniques for gene cloning and gene insertion, we are now entering an era of genetic exploitation of other organisms undreamed of only 50 years ago. We can now begin to design genes to produce in other organisms novel gene products for the benefit of human beings; we are no longer restricted to selecting useful genes that arise by mutation. We are, however, only at the beginning of this new era, and so far we have only scratched the surface of the knowledge that is required for true engineering and design of protein molecules. We distinguish **protein engineering**, by which we mean mutating the gene of an existing protein in an attempt to alter its function in a *predictable* way, from **protein design**, which has the more ambitious goal of designing *de novo* a protein to fulfill a desired function.

Protein engineers frequently have been surprised by the range of effects caused by single mutations that they hoped would change only one specific and simple property in enzymes; some examples are described in Chapter 15. The often surprising results of such experiments reveal how little we know about the rules of protein stability and the energetics of ligand binding and catalytic efficiency; they also serve to emphasize how difficult it is to design *de novo* stable proteins with specific functions. However, by using the methods of engineering and design, we are now at least increasing rapidly our basic knowledge of the function of protein molecules. For example, we now know that the difference in energetic terms between the transition states of a naturally evolved useful enzyme and an engineered useless mutant corresponds to less than the energy of a single hydrogen bond, even for such important life-sustaining enzymes as the CO₂-fixing enzyme in green plants, rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase).

Knowledge of a protein's tertiary structure is a prerequisite for the proper engineering of its function. Unfortunately, in spite of recent significant techno-

Exhibit 3

This paper was presented at the National Academy of Sciences colloquium "Proteolytic Processing and Physiological Regulation," held February 20-21, 1999, at the Arnold and Mabel Beckman Center in Irvine, CA.

Reverse biochemistry: Use of macromolecular protease inhibitors to dissect complex biological processes and identify a membrane-type serine protease in epithelial cancer and normal tissue

TOSHIHIKO TAKEUCHI*, MARC A. SHUMAN†, AND CHARLES S. CRAIK*‡

*Departments of Pharmaceutical Chemistry and Biochemistry & Biophysics, and †Department of Medicine, University of California, San Francisco, CA 94143

ABSTRACT Serine proteases of the chymotrypsin fold are of great interest because they provide detailed understanding of their enzymatic properties and their proposed role in a number of physiological and pathological processes. We have been developing the macromolecular inhibitor ecotin to be a "fold-specific" inhibitor that is selective for members of the chymotrypsin-fold class of proteases. Inhibition of protease activity through the use of wild-type and engineered ecotins results in inhibition of rat prostate differentiation and retardation of the growth of human PC-3 prostatic cancer tumors. In an effort to identify the proteases that may be involved in these processes, reverse transcription-PCR with PC-3 poly(A)⁺ mRNA was performed by using degenerate oligonucleotide primers. These primers were designed by using conserved protein sequences unique to chymotrypsin-fold serine proteases. Five proteases were identified: urokinase-type plasminogen activator, factor XII, protein C, trypsinogen IV, and a protease that we refer to as membrane-type serine protease 1 (MT-SP1). The cloning and characterization of the MT-SP1 cDNA shows that it encodes a mosaic protein that contains a transmembrane signal anchor, two CUB domains, four LDLR repeats, and a serine protease domain. Northern blotting shows broad expression of MT-SP1 in a variety of epithelial tissues with high levels of expression in the human gastrointestinal tract and the prostate. A His-tagged fusion of the MT-SP1 protease domain was expressed in *Escherichia coli*, purified, and autoactivated. Ecotin and variant ecotins are subnanomolar inhibitors of the MT-SP1 activated protease domain, suggesting a possible role for MT-SP1 in prostate differentiation and the growth of prostatic carcinomas.

Serine proteases possessing a chymotrypsin fold are of great interest because they provide detailed understanding of their enzymatic properties and their proposed role in a number of physiological and pathological processes. A wealth of information exists on structure-function relationships regarding this large class of enzymes. Moreover, potent and specific inhibitors are readily available for use in dissecting the function of these enzymes. These proteases exist as precursors that are activated by specific and limited proteolysis, allowing regulation of enzyme activity (1). Examples of this type of regulation include blood coagulation (2), fibrinolysis (3), complement activation (4), and trypsinogen activation by enteropeptidase in digestion (5). The precise control of these activation processes is crucial for normal physiological enzymatic function; misregulation of these enzymes can lead to pathological conditions (2-5).

We are interested in studying the role of these chymotrypsin-fold serine proteases in cancer by using a "fold-specific"

inhibitor, ecotin (6, 7). Ecotin or engineered versions of ecotin can be introduced into complex biological systems as probes of proteolysis by these chymotrypsin-fold proteases. If effects are observed on treatment with these unique inhibitors, then the large body of knowledge concerning the biochemistry of these proteases can be tapped to understand the structure and function of the target proteases. For example, the molecular cloning, structural modeling, and mechanistic understanding of the enzymes are immediately accessible. We refer to this approach, which is analogous to "reverse genetics," as "reverse biochemistry," and we have applied it to identification of specific serine proteases in prostate cancer.

Urokinase-type plasminogen activator (uPA) has been implicated in tumor-cell invasion and metastasis. Cancer-cell invasion into normal tissue can be facilitated by uPA through its activation of plasminogen, which degrades the basement membrane and extracellular matrix (reviewed in refs. 8 and 9). The role of other serine proteases in cancer has been less well characterized.

One useful model system for studying many issues that are pertinent to prostate cancer is the development of the rodent ventral prostate in explant cultures. Macromolecular inhibitors of serine proteases of the chymotrypsin fold, ecotin and ecotin M84R/M85R (6, 7), inhibit ductal branching morphogenesis and differentiation of the explanted rat ventral prostate (F. Elfmann, T.T., C.C., G. Cunha, and M.S., unpublished data). Ecotin M84R/M85R is a 2,800-fold more potent inhibitor of uPA than ecotin (1 nM vs. 2.8 μ M) (6). However, inhibition of prostate differentiation was seen with both inhibitors, suggesting that uPA and other related serine proteases are involved in the differentiation and continued growth of the rat ventral prostate. Thus, unidentified serine proteases may play a role in growth and prevention of apoptosis in prostate epithelial cells in this system.

Another well characterized model that is derived from human prostate cancer epithelial cells is the PC-3 cell line (10). The PC-3 cell line expresses uPA as assayed by ELISA and by Northern blotting of PC-3 mRNA (11). We found that the primary tumor size in PC-3-implanted nude mice was significantly smaller in both ecotin M84R/M85R and ecotin wild-type treated mice treated for 7 weeks compared with the primary tumor size of PBS-treated mice. Metastasis from the primary tumors were similarly lower in the inhibitor-treated

Abbreviations: MT-SP1, membrane-type serine protease 1; CUB, complement factor 1R-urchin embryonic growth factor-bone morphogenetic protein; LDLR, low density lipoprotein receptor; uPA, urokinase-type plasminogen activator; pNA, *p*-nitroanilide. Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. BankIt257050 and AF133086).

‡To whom reprint requests should be addressed. E-mail: craik@cgl.ucsf.edu.

mice than in PBS-treated mice (O. Melnyk, T.T., C.C., and M.S., unpublished data). Inhibition was not unexpected with ecotin M84R/M85R treatment, because uPA has been implicated in metastasis. However, wild-type ecotin is a poor, micromolar inhibitor of uPA; one interpretation of the data is that the decrease in tumor size and metastasis in the mouse model involves the inhibition of additional serine proteases. Thus, identification of the serine proteases expressed by PC-3 prostate cells may provide insight into the role of these proteases in cancer and prostate growth and development. In this report we have extended the strategy of using PCR with degenerate oligonucleotide primers that were designed by using conserved sequence homology (12–14) to identify additional serine proteases made by cancer cells. Five independent serine protease cDNAs derived from PC-3 mRNA were sequenced, including a novel serine protease, which we refer to as membrane-type serine protease 1 (MT-SP1), and the cloning and characterization of this cDNA that encodes a mosaic, transmembrane protease is reported.

MATERIALS AND METHODS

Materials. All primers used were synthesized on a Applied Biosystems 391 DNA synthesizer. All restriction enzymes were purchased from New England Biolabs. Automated DNA sequencing was carried out on an Applied Biosystems 377 Prism sequencer, and manual DNA sequencing was carried out under standard conditions. N-terminal amino acid sequencing was performed on an ABI 477A by the University of California, San Francisco Biomolecular Resource Center. The synthetic substrates, Suc-AAPX-*p*-nitroanilide (pNA), [*N*-succinyl-alanyl-alanyl-prolyl-Xxx-pNA (Xxx = alanyl, aspartyl, glutamyl, phenylalanyl, leucyl, methionyl, or arginyl)], and H-Arg-pNA, (arginyl-pNA), were purchased from Bachem. Deglycosylation was performed by using PNGase F (NEB, Beverly, MA). All other reagents were of the highest quality available and purchased from Sigma or Fisher unless otherwise noted.

Isolation of cDNA from PC-3 Cells. mRNA was isolated from PC-3 cells by using the polyATtract System 1000 kit (Promega). Reverse transcription was primed by using the "lock-docking" oligo(dT) primer (15). Superscript II reverse transcriptase (Life Technologies, Grand Island, NY) was used in accordance with the manufacturer's instructions to synthesize the cDNA from the PC-3 mRNA.

Amplification of MT-SP1 Gene. The degenerate primers used for amplifying the protease domains were designed from the consensus sequences flanking the catalytic histidine (5' His-primer) and the catalytic serine (3' Ser-primer), similar to those described (12). The 5' primer used is as follows: 5'-TGG (AG)TI (CAG)TI (AT)(GC)I GCI (GA)CI CA(CT) TG-3', where nucleotides in parentheses represent equimolar mixtures and I represents deoxyinosine. This primer encodes at least the following amino acid sequence: W (I/V) (I/V/L/M) (S/T) A (A/T) H C. The 3' primer used is as follows: 5'-IGG ICC ICC I(GC)(AT) (AG)TC ICC (CT)TI (GA)CA IG(ATC) (GA)TC-3'. The reverse complement of the 3' primer encodes at least the following amino acid sequence: D (A/S/T) C (K/E/Q/H) G D S G G P.

Direct amplification of serine protease cDNA was not possible by using the above primers. Instead, the first PCR was performed with the 5' His-primer and the oligo(dT) primer described above, by using the "touchdown" PCR protocol (16), with annealing temperatures decreasing from 52°C to 42°C over 22 rounds and 13 final rounds at 54°C annealing temperature. Cycle times were 1 min (denaturing), 1 min (annealing), and 2 min (extension) and were followed by one final extension time of 15 min after the final round of PCR. The template for the second PCR was 0.5 μ L (total reaction volume 50 μ L) of a 1:10 dilution of the first PCR mixture that was performed

with the 5' His-primer and the oligo(dT). The second PCR reaction was primed with the 5' His- and the 3' Ser-primers and performed by using the touchdown protocol described above. All PCRs used 12.5 pmol of primer for 50- μ L reaction volume.

The product of the second reaction was purified on a 2% agarose gel, and all products between 400 and 550 bp were cut from the gel and extracted by using the QIAquick gel extraction kit (Qiagen, Chatsworth, CA). These products were digested with the *Bam*HI restriction enzyme to cut any uPA cDNA, and all 400- to 500-bp fragments were repurified on a 2% agarose gel. These reaction products were subjected to a third PCR by using the 5' His-primer and the 3' Ser-primer by using the identical touchdown procedure. These reaction products were gel-purified and directly cloned into the pPCR2.1 vector by using the TOPO TA ligation kit (Invitrogen). DNA sequencing of the inserts determined the cDNA sequence from nucleotides 1,984 to 2,460 (see Fig. 1).

Northern Blot Analysis. ³²P-labeled nucleotides were purchased from Amersham Pharmacia. A cDNA fragment containing nucleotides 1,173–2,510 was digested from expressed sequence tag w39209 by using restriction enzymes *Eco*RI and *Bsm*BI, yielding a 1.3-kilobase nucleotide insert. Labeled cDNA probes were synthesized by using the Rediprime random primer labeling kit (Amersham Pharmacia) and 20 ng of the purified insert. Poly(A)+ RNA membranes for Northern blotting were purchased from Origene (Rockville, MD; HB-1002, HB-1018) and CLONTECH (Human II 7759–1, Human Cancer Cell Line 7757). The blots were performed under stringent annealing conditions as described in ref. 17.

Construction of Expression Vectors. The mature protease domain and a small portion of the pro-domain (nucleotides 1,822–2,601) cDNA were amplified by using PCR from expressed sequence tag w39209 and ligated into the pQE30 vector (Qiagen). This construct is designed to overexpress the protease sequence from amino acids (aa) 596–855 with the following fusion: Met-Arg-Gly-Ser-His-aa596–855. The His-tag fusion allows affinity purification by using metal-chelate chromatography. The change from Ser-805, encoded by TCC, to Ala (GCT) was performed by using PCR. The presence of the correct Ser → Ala substitution in the pQE30 vector was verified by DNA sequence analysis.

Expression and Purification of the Protease Domain. The above-mentioned plasmids were separately transformed into *Escherichia coli* X-90 to afford high-level expression of recombinant protease gene products (18). Expression and purification of the recombinant enzyme from solubilized inclusion bodies was performed as described (19). Protein-containing fractions were pooled and dialyzed overnight at 4°C against 50 mM Tris (pH 8), 10% glycerol, 1 mM 2-mercaptoethanol, and 3 M urea. Autoactivation of the protease was monitored on dialysis against storage buffer (50 mM Tris, pH 8/10% glycerol) at 4°C by using the substrate Spectrozyme tPA (hexahydroxyrosyl-Gly-Arg-pNA, American Diagnostica, Greenwich, CT). Hydrolysis of Spectrozyme tPA was monitored at 405 nm for the formation of *p*-nitroaniline by using a Uvikon 860 spectrophotometer. Activated protease was bound to an immobilized *p*-aminobenzamidine resin (Pierce) that had been equilibrated with storage buffer. Bound protease was eluted with 100 mM benzamidine and the protein containing fractions were pooled. Excess benzamidine was removed by using FPLC with a Superdex 70 (Amersham Pharmacia) gel filtration column that was equilibrated with storage buffer. Protein containing fractions were pooled and stored at –80°C. The cleavage of the purified Ser⁸⁰⁵Ala protease domain was performed at 37°C by addition of active recombinant protease domain to 10 nM. Cleavage was monitored by using SDS/PAGE.

Determination of Substrate Kinetics. The purified serine protease domain was titrated with 4-methylumbelliferyl *p*-guanidinobenzoate (MUGB) to obtain an accurate concen-

FIG. 1. Nucleotide sequence of the cDNA encoding human MT-SP1 and predicted protein sequence. Numbering indicates nucleotide or amino acid residue. Amino acids are shown in single-letter code. The termination codon is shown by *. The underlined stop codon at nucleotide 10 is in frame with the initiating methionine. The Kozak consensus sequence (24) at the start codon is underlined at nucleotide 32. The predicted *N*-glycosylation sites at amino acids 109, 302, 485, and 772 are underlined. A possible polyadenylation sequence (46) at nucleotide 3,120 is also underlined. The catalytic triad in the serine protease domain is highlighted: His-656, Asp-711, and Ser-805.

Inhibition of MT-SPI Protease Domain with Ecotin and Ecotin M84R/M85R. Ecotin and ecotin M84R/M85R were purified from *E. coli* as described (6). Various concentrations of ecotin or ecotin M84R/M85R were incubated with the His-tagged serine protease domain in a total volume of 990 μ l of buffer containing 50 mM NaCl, 50 mM Tris-HCl (pH 8.8), and 0.01% Tween 20. Ten microliters of Spectrozyme tPA was added, yielding a solution containing 100 μ M substrate. The final enzyme concentration was 63 pM, and the ecotin and ecotin M84R/M85R concentration ranged from 0.1 to 50 nM. The data were fit to the equation derived for kinetics of reversible tight-binding inhibitors (21, 22), and the values for apparent K_i were determined.

Cloning of Serine Protease Domain cDNAs from PC-3 Cells and Amplification of MT-SP1 cDNA. PCR amplification of serine protease cDNA was performed by using "consensus

cloning", where the amplification was performed with degenerate primers designed to anneal to cDNA encoding the region about the conserved catalytic histidine (5' His-primer) and the conserved catalytic serine (3' Ser-primer). The consensus primers were designed by using 37 human sequences within a sequence alignment of 242 serine proteases of the chymotrypsin fold that are reported in the SwissProt database. To bias the screen for previously unidentified proteases in the PC-3 cDNA, uPA cDNA was cut and removed by using the known *Bam*HI endonuclease site in the uPA cDNA sequence. The expected size of the cDNA fragments amplified between His-57 and Ser-195 cDNA (standard chymotrypsinogen numbering) is between 400 and 550 bp; statistically, only 1 in 10 cDNAs of that length will be cleaved by *Bam*HI. Thus, cDNAs obtained from the PCR reactions with the 5' His-primer and 3' Ser-primer were size selected for the 400- to 550-bp range, digested with *Bam*HI, and purified from any digested cDNAs. After a subsequent round of PCR, the products were cloned into pPCR2.1 (Fig. 2). Twenty clones were digested with *Eco*RI to monitor the size of the cDNA insert. Three clones lacked inserts of the correct size. The remaining 17 clones containing inserts between 400 and 550 bp were sequenced. BLAST searches of the resulting sequences revealed that six clones did not match serine protease sequences. The remaining cDNAs yielded clones corresponding to factor XII (two clones), protein C (two clones), trypsinogen type IV (two clones), uPA (one clone), and MT-SP1 (four clones). Additional serine protease sequences may not have been found because they were digested by *Bam*HI, lost in the size selection, or present in lower frequencies.

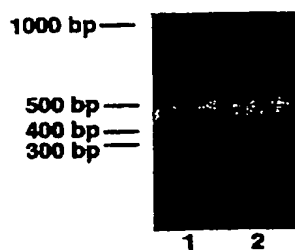


FIG. 2. Lane 1 shows the PCR products obtained by using degenerate primers designed from the consensus sequences flanking the catalytic histidine (5' His-primer) and the catalytic serine (3' Ser-primer). The products remaining between 400 and 550 bp after digestion with *Bam*HI were reamplified by using the same degenerate primers. The products from this second PCR are shown in Lane 2.

Multiple expressed sequence tag sequences were found for the cDNA. Expressed sequence tag accessions aa459076, aa219372, and w39209 were used extensively for sequencing the cDNA starting from nucleotide 746 and 2,461–3,142, but no start codon was observed. A sequence was also found in GenBank (accession no. U20428). This sequence also lacks the 5' end of the cDNA but allowed amplification of cDNA from nucleotides 196–745. Rapid amplification of cDNA ends (RACE) (23) was used to obtain further 5' cDNA sequence. Application of RACE did not yield a clone containing the entire 5'-untranslated region, but the sequence obtained contained a stop codon in-frame with the Kozak start sequence (24), giving confidence that the full coding sequence of the cDNA has been obtained. The nucleotide sequence and predicted amino acid sequence are shown in Fig. 1.

The nucleotide sequence surrounding the proposed start codon matches the optimal sequence of ACCATGG for translation initiation sites proposed by Kozak (24). In addition, there is a stop codon in-frame with the putative start codon, which gives further evidence that initiation occurs at that site. The DNA sequence predicts an 855-aa mosaic protein composed of multiple domains (Fig. 3). The coding sequence does not contain a typical signal peptide but does contain a single

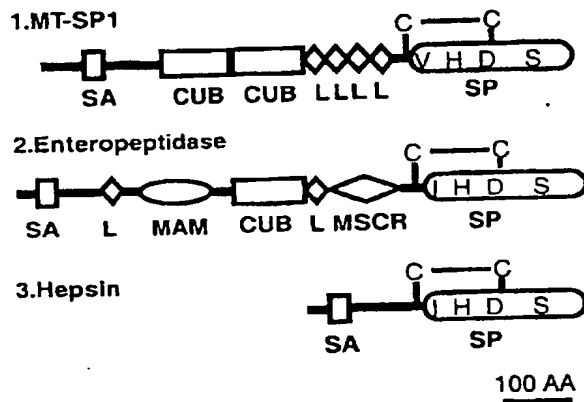
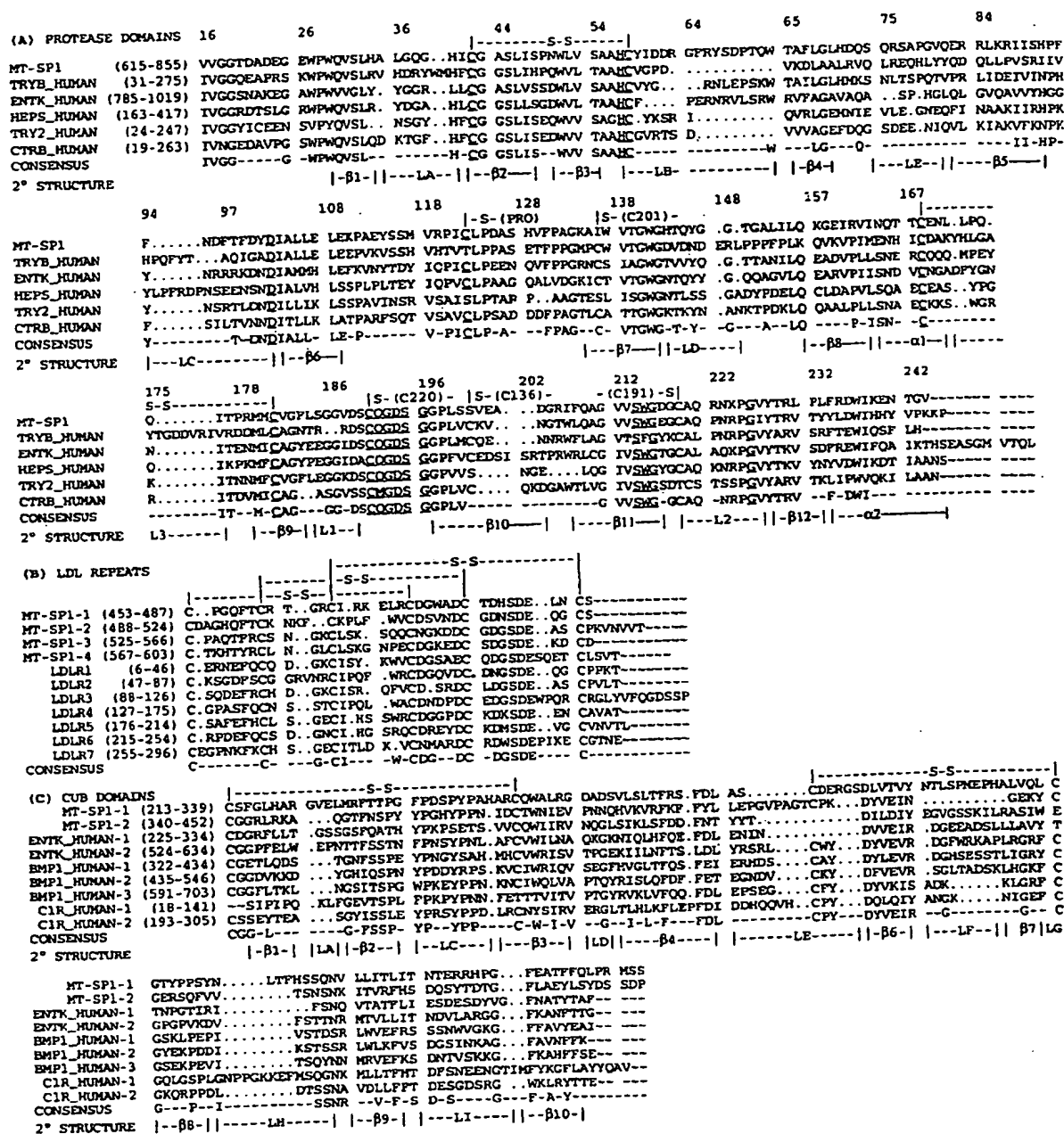


FIG. 3. The domain structure of human MT-SP1 is compared with the domain structure of enteropeptidase (47) and hepsin (25). SA, possible signal anchor; CUB, a repeat first identified in complement components C1r and C1s, the urchin embryonic growth factor and bone morphogenetic protein 1 (27); L, LDLR repeat (29); SP, a chymotrypsin family serine protease domain (40); MAM, a domain homologous to members of a family defined by meprin, protein A5, and the protein tyrosine phosphatase μ (48); MSCR, a macrophage scavenger receptor cysteine-rich motif (29). The predicted disulfide linkages are shown labeled as C-C.

hydrophobic sequence of 26 residues (residues 55–81), which is flanked by a charged residue on each side. This sequence may constitute a signal anchor sequence, similar to that observed in other proteases, including hepsin (25) and enteropeptidase (26). Following the putative signal anchor sequence are two complement factor 1R-urchin embryonic growth factor-bone morphogenetic protein (CUB) domains (27), which are named after the proteins in which the modules were first discovered: complement subcomponents C1s and C1r, urchin embryonic growth factor (Uegf), and bone morphogenetic protein 1 (BMP1). CUB domains have conserved characteristics, which include the presence of four cysteine residues and various conserved hydrophobic and aromatic positions (27). The CUB domain, which has recently been characterized crystallographically (28), consists of 10 β -strands that are organized into two 5-stranded β -sheets. Following the CUB domains are four low-density lipoprotein receptor (LDLR) repeats (29), which are named after the receptor ligand-binding repeats that are present in the LDLR. These repeats have a highly conserved pattern and spacing of six cysteine residues that form three intramolecular disulfide bonds. The final domain observed is the serine protease domain. The alignments of these domains with other members of their respective classes are shown in Fig. 4.

Tissue Distribution of MT-SP1 mRNA. Northern blots of human poly(A)⁺ RNA, made by using a 1.3-kilobase fragment of MT-SP1 cDNA fragment as a probe, show a ~3.3-kilobase fragment appearing in epithelial tissues including the prostate, kidney, lung, small intestine, stomach, colon, and placenta, as well as other tissues, including spleen, liver, leukocytes, and thymus. This band was not observed in muscle, brain, ovary, or testis (Fig. 5). Similar experiments performed on a human cancer cell line blot shows that MT-SP1 is expressed in the colorectal adenocarcinoma, SW480, but was not observed in the promyelocytic leukemia HL-60, HeLa cell S3, chronic myelogenous leukemia K-562, lymphoblastic leukemia MOLT-4, Burkitt's lymphoma Raji, lung carcinoma A549, or melanoma G361 lanes (data not shown). This 3.3-kilobase mRNA fragment is slightly longer than the 3.1-kilobase sequence presented in Fig. 5, suggesting that there may still be sequence in the 5'-untranslated region that has not been identified.

Activation and Purification of His-MT-SP1 Protease Domain. The serine protease domain of MT-SP1 was expressed in *E. coli* as a His-tagged fusion and was purified from inclusion bodies under denaturing conditions by using metal-chelate affinity chromatography. The yield of enzyme after this step was ~3 mg of protein per liter of *E. coli* culture. This denatured protein refolded when the urea was dialyzed from the protein. Surprisingly, the purified renatured protein showed a time-dependent shift on an SDS/PAGE gel (Fig. 6A), with the lower fragment being the size of the mature, processed enzyme lacking the His tag. N-terminal sequencing of the purified, activated protease domain yielded the expected VVGTT activation sequence. When the refolded protein was tested for activity by using the synthetic substrate Spectrozyme tPA, a time-dependent increase in activity was observed (Fig. 6B). In contrast, the protease domain that contains the Ser⁸⁰⁵Ala mutation showed neither a change in size on an SDS polyacrylamide gel nor an increase in enzymatic activity under identical conditions (data not shown), suggesting that the catalytic serine is necessary for activation and is not the result of a contaminating protease. To show that the cleavage of the protease domain was a result of His-tagged MT-SP1 protease activity, the inactive Ser⁸⁰⁵Ala protease domain was treated with purified recombinant enzyme (Fig. 6C). This treatment results in the formation of a cleavage product that corresponds to the size of the active protease (Fig. 6C, lane 7). Untreated protease domain does not get cleaved (Fig. 6C, lane 8). From these results, it is concluded that the protease autoactivates on



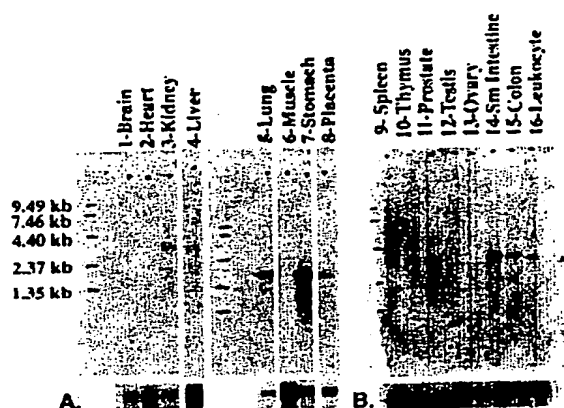


FIG. 5. Tissue distribution of MT-SP1 mRNA levels. Northern blots of human poly(A)⁺ RNA from assorted human tissues was hybridized with radiolabeled cDNA probes as described in *Materials and Methods*. Upper shows hybridization by using a MT-SP1 1.3-kilobase cDNA fragment derived from expressed sequence tag clone w39209 and exposed overnight. Lower shows the same blot after being stripped and rehybridized with a loading standard β -actin (A) or human glyceraldehyde phosphate dehydrogenase (GAPDH) (B) cDNA probe exposed for 2 hours. The mobility of RNA size standards is indicated at the left.

Purified protease domain was tested for hydrolytic activity against tetrapeptide substrates of the form Suc-AAPX-pNA, which contained various amino acids at the P1 position (P1-Ala, Asp, Glu, Phe, Leu, Met, Lys, or Arg). The only substrates with detectable activity were those with P1-Lys or P1-Arg. The serine protease domain with the Ser⁸⁰⁵Ala mutation had no detectable activity. The activity of the protease domain was further characterized by using the substrate Spectrozyme tPA, yielding: $K_m = 31.4 \pm 4.2 \mu\text{M}$, $k_{cat} = 2.6 \times 10^2 \pm 6.5 \text{ s}^{-1}$, and $k_{cat}/K_m = 6.9 \times 10^6 \pm 2.3 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$. Ecotin inhibition of the MT-SP1 His-tagged protease domain fits a tight-binding reversible inhibitory model (21, 22) as observed for ecotin interaction with other serine protease targets (6, 7, 30). Inhibition assays by using ecotin and ecotin M84R/M85R yielded apparent K_i values of $782 \pm 92 \text{ pM}$ and $9.8 \pm 1.5 \text{ pM}$, respectively.

DISCUSSION

Structural Motifs of MT-SP1. In this work, we characterize the expression of chymotrypsin-fold proteases by PC-3 cells and cloned a member of this family we call MT-SP1. The name membrane-type serine protease 1 (MT-SP1) is given to be consistent with the nomenclature of the membrane-type matrix metalloproteases (MT-MMPs; ref. 32). The cDNA likely encodes a membrane-type protein because of the lack of a signal sequence and the presence of a putative SA that is also seen in other membrane-type serine proteases hepsin (25), enteropeptidase (26), and TMPRSS2 (32), and human airway trypsin-like protease (33). We propose that proteins that are localized to the membrane through a SA and that encode a chymotrypsin fold serine protease domain be categorized in the MT-SP family. The membrane localization of MT-SP1 is supported by immunofluorescence experiments that localize the protease domain to the extracellular cell surface (unpublished results).

Following the putative SA are several domains that are thought to be involved in protein-protein interactions or protein-ligand interactions. For example, CUB domains can mediate protein-protein interactions as with the seminal plasma PSP-I/PSP-II heterodimer that is built by CUB-domain interactions (28) and with procollagen C-proteinase

enhancer protein and procollagen C-proteinase (BMP-1) (34, 35). Interestingly, most of the proteins that contain CUB domains are involved in developmental processes or are involved in proteolytic cascades (27), which suggests that MT-SP1 may play a similar role. The four repeated motifs that follow the CUB domains are known as LDLR ligand-binding repeats, named after the seven copies of repeats found in the LDLR. There are several negatively charged amino acids between the fourth and sixth cysteines that are highly conserved in the LDLR and are also seen in the LDLR repeats of MT-SP1. The conserved motif Ser-Asp-Glu (residues 44–46 in Fig. 4) are known to be important for binding the positively charged residues of the LDLR ligands apolipoprotein B-100 (ApoB-100) and ApoE (29). The ligand-binding repeats of MT-SP1 most likely do not mediate interaction with ApoB-100 or ApoE but may be involved in the interaction with other positively charged ligands. For example, LDLR repeats in the LDLR-related protein have been implicated the binding and recycling of protease-inhibitor complexes such as uPA-plasminogen activator inhibitor-1 (PAI-1) complexes (reviewed in refs. 36 and 37). It also has been shown that the pro domain of enteropeptidase is involved in interactions with its substrate trypsinogen, allowing 520-fold greater catalytic efficiency in the cleavage compared with the protease domain alone (38). By analogy, similar interactions should occur between MT-SP1 and its substrates. Thus, further investigation of MT-SP1 CUB domain or LDLR repeat interactions may yield insight into the function of this protein.

The amino acid sequence of the serine protease domain of MT-SP1 is highly homologous to other proteases found in the family (Fig. 4). The essential features of a functional serine protease are contained in the deduced amino acid sequence of

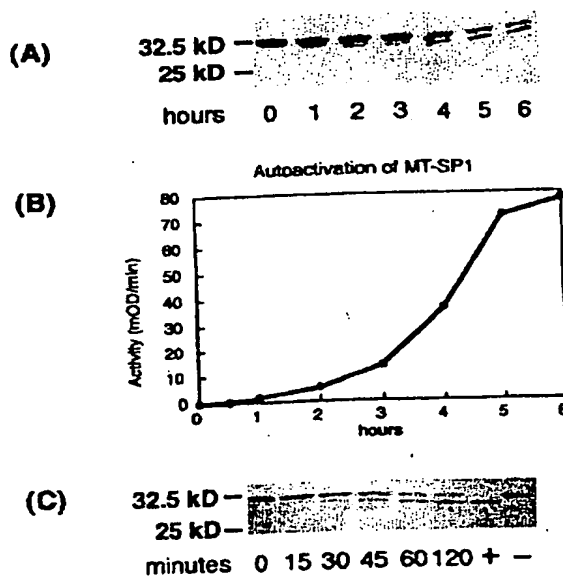


FIG. 6. Activation and purification of His-tagged MT-SP1 protease domain. A representative experiment is shown in A and B. (A) Activation at 4°C was monitored by using SDS/PAGE. The upper band represents inactivated protease domain, and the lower band represents active protease (also verified by N-terminal sequencing). (B) The activation of the protein was monitored by using Spectrozyme tPA as a synthetic substrate for the protease domain. (C) Inactive Ser⁸⁰⁵Ala protease domain is cleaved with 10 nM activated His-tagged MT-SP1 protease domain at 37°C. The specific cleavage of active MT-SP1 protease domain is required for proper processing at the activation site. Active protease domain is shown in lane 7 (+), and no cleavage of the untreated inactive protease domain is observed (lane 8, -).

the domain. The residues that comprise the catalytic triad, His-656, Asp-711, and Ser-805, corresponding to His-57, Asp-102, and Ser-195 in chymotrypsin, are observed in MT-SP1 (for reviews, see refs. 39 and 40). The sequence Ser²¹⁴Trp²¹⁵Gly²¹⁶ (Ser⁸²⁵Trp⁸²⁶Gly⁸²⁷), which is thought to interact with the side chains of the substrate for properly orienting the scissile bond is present. Gly-193 (Gly-803) and Gly-196 (Gly-805), which are thought to be necessary for proper orientation of Ser-195 (Ser-805), also are present. Based on homology to chymotrypsin, three disulfide bonds are predicted to form within the protease domain at Cys-44–Cys-58, Cys-168–Cys-182, and Cys-191–Cys-220 (Cys-643–Cys-657, Cys-776–Cys-790, and Cys-801–Cys-830), and a fourth disulfide bond should form between the catalytic and the pro-domain Cys-122–Cys-1 (Cys-731–Cys-604), as observed for chymotrypsin. This predicted disulfide with the pro domain suggests that the active catalytic domain should still be localized to the cell surface via a disulfide linkage. The presence of the catalytic machinery and other conserved structural components described above suggest that all features necessary for proteolytic activity are present in the encoded sequence.

Substrate Specificity of the MT-SP1 Protease Domain. The S1 site specificity (41) of a protease is largely determined by the amino acid residue at position 189. This position is occupied by an aspartate in MT-SP1, suggesting that the protease has specificity for Arg/Lys in the P1 position. In addition, the presence of a polar Gln-192 (Gln-803), as in trypsin, is consistent with basic specificity. Furthermore, the presence of Gly-216 (Gly-827) and Gly-226 (Gly-837) is consistent with the presence of a deep S1 pocket, unlike elastase, which has Val-216 and Thr-226 that block the pocket and thereby contribute to the P1 specificity for small hydrophobic side chains. The specificity at the other subsites is largely dependent on the nature of the seven loops A–E and loops 2 and 3 (Fig. 4). Loop C in enterokinase has a number of positively charged residues that are thought to interact with the negatively charged activation site in trypsinogen, Asp-Asp-Asp-Asp-Lys (26). One known substrate for MT-SP1 (as described below) is the activation site of MT-SP1, which is Arg-Gln-Ala-Arg (residues 611–614). Loop C contains two Asp residues that may participate in the recognition of the activation sequence.

One means of obtaining further data on substrate specificity is by characterization of the activity of the recombinant proteolytic domain. Enterokinase has been characterized from both recombinant (38, 42) and native (43, 44) sources. However, proteolytic activity for the other reported membrane-type serine proteases hepsin (25) and TMPRSS2 (32) are only predicted based on sequence homology. To produce active recombinant MT-SP1, a His-tagged fusion of the protease domain was cloned into an *E. coli* vector and expressed and purified to homogeneity. Fortunately, the protease domain refolded and autoactivated after resuspension and purification from inclusion bodies. This activity, coupled with the lack of activity in the Ser¹⁹⁵Ala (Ser⁸⁰⁵Ala) variant, demonstrates that the cDNA encodes a catalytically proficient protease. Autoactivation of the protease domain at the arginine-valine site (Arg⁶¹⁴-Val⁶¹⁵) shows that the protease has Arg/Lys specificity as predicted by the sequence homology to other proteases of basic specificity. Specificity and selectivity are confirmed by the lack of cleavage of AAPX-pNA substrates that do not have x = R, K. Further characterization with Spectrozyme tPA revealed an active enzyme with $k_{cat} = 2.6 \times 10^2 \text{ s}^{-1}$. However, the His-tagged serine protease domain does not cleave H-Arg-pNA, showing that, unlike trypsin, there is a requirement for additional subsite occupation for catalytic activity. This suggests that the enzyme is involved in a regulatory role that requires selective processing of particular substrates rather than nonselective degradation.

MT-SP1 Function. In other studies, we have found that inhibition of serine protease activity by ecotin or ecotin

M84R/M85R inhibits testosterone-induced branching ductal morphogenesis and enhances apoptosis in a rat ventral prostate model (F. Elfman, T.T., C.S.C., G. Cunha, and M.A.S., unpublished results). Moreover, the rat homolog of MT-SP1 is expressed in the normal rat ventral prostate (data not shown). Assays of the protease domain with ecotin and ecotin M84R/M85R showed that the enzymatic activity is strongly inhibited ($782 \pm 92 \text{ pM}$ and $9.8 \pm 1.5 \text{ pM}$, respectively), suggesting that rat MT-SP1 is likely to be inhibited at the concentrations of these inhibitors used in our experiments. MT-SP1 inhibition may result in the observed inhibition of differentiation and/or increased apoptosis. Future studies are aimed at definitively resolving the role of MT-SP1 in prostate differentiation. The broad expression of MT-SP1 in epithelial tissues is consistent with the possibility that it is involved in cell maintenance or growth, perhaps by activating growth factors or by processing prohormones.

MT-SP1 may participate in a proteolytic cascade that results in cell growth and/or differentiation. Another structurally similar membrane-type serine protease, enteropeptidase (Fig. 3), is involved in a proteolytic cascade by which activation of trypsinogen leads to activation of downstream intestinal proteases (5). Enteropeptidase is expressed only in the enterocytes of the proximal small intestine, thus precisely restricting activation of trypsinogen. Thus, in contrast to secreted proteases that may diffuse throughout the organism, the membrane association of MT-SP1 should also allow the proteolytic activity to be precisely localized, which may be important for proper physiological function; improper localization of the enzyme, or levels of downstream substrates could lead to disease.

We have found subcutaneous coinjection of PC-3 cells with wild-type ecotin or ecotin M84R/M85R led to a decrease in the primary tumor size compared with animals in whom PC-3 cells and saline were injected (O. Melnyk, T.T., C.S.C. and, M.A.S., unpublished results). Because wild-type ecotin is a poor, micromolar inhibitor of uPA, serine proteases other than uPA likely are involved in this primary tumor proliferation. Both wild-type ecotin and ecotin M84R/M85R are potent, subnanomolar inhibitors of MT-SP1, raising the possibility that MT-SP1 plays an important role in progression of epithelial cancers expressing this protease.

Direct biochemical isolation of the substrates may be possible if MT-SP1 adhesive domains such as the CUB domains or LDLR repeats interact with the substrates. In addition, likely substrates may be predicted and tested for by using knowledge of extended enzyme specificity. For example, the characterization of the substrate specificity of granzyme B allowed the prediction and confirmation of substrates for this serine protease (45). Thus, these complimentary studies should further shed light on the physiological function of this enzyme.

We thank Marion Conn, Robert Maeda, Todd Pray, Ibrahim Adiguzel, and Ralph Reid for technical assistance and helpful discussions. T.T. was supported by a National Institutes of Health postdoctoral fellowship CA71097, and this work was supported by National Institutes of Health Grant CA72006.

1. Neurath, H. & Walsh, K. A. (1976) *Proc. Natl. Acad. Sci. USA* 73, 3825–3832.
2. Davie, E. W., Fujikawa, K. & Kisiel, W. (1991) *Biochemistry* 30, 10363–10370.
3. Chandler, W. L. (1996) *Crit. Rev. Oncol. Hematol.* 24, 27–45.
4. Reid, K. B. M. & Porter, R. R. (1981) *Annu. Rev. Biochem.* 50, 433–464.
5. Huber, R. & Bode, W. (1978) *Acc. Chem. Res.* 11, 114–122.
6. Wang, C.-I., Yang, Q. & Craik, C. S. (1995) *J. Biol. Chem.* 270, 12250–12256.
7. Yang, S. Q., Wang, C.-I., Gillmor, S. A., Fletcher, R. J. & Craik, C. S. (1998) *J. Mol. Biol.* 279, 945–957.

8. Dano, K., Andreassen, P. A., Grondahl-Hansen, J., Kristensen, P., Nielsen, L. S. & Skriver, L. (1985) *Adv. Cancer Res.* 44, 139–266.
9. Andreassen, P. A., Kjoller, L., Christensen, L. & Duffy, M. J. (1997) *Int. J. Cancer* 72, 1–22.
10. Kaighn, M. E., Narayan, K. S., Ohnuki, Y., Lechner, J. F. & Jones, L. W. (1979) *Invest. Urol.* 17, 16–23.
11. Yoshida, E., Verrusio, E. N., Mihara, H., Oh, D. & Kwaan, H. C. (1994) *Cancer Res.* 54, 3300–3304.
12. Sakanari, J. A., Staunton, C. E., Eakin, A. E., Craik, C. S. & McKerrow, J. H. (1989) *Proc. Natl. Acad. Sci. USA* 86, 4863–4867.
13. Wiegand, U., Corbach, S., Minn, A., Kang, J. & Muller-Hill, B. (1993) *Gene* 136, 167–175.
14. Kang, J., Wiegand, U. & Muller-Hill, B. (1992) *Gene* 110, 181–187.
15. Borson, N. D., Salo, W. L. & Drewes, L. R. (1992) *PCR Methods Appl.* 2, 144–148.
16. Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. & Mattick, J. S. (1991) *Nucleic Acids Res.* 19, 4008.
17. Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K., eds. (1990) *Current Protocols in Molecular Biology* (Wiley, New York).
18. Evnin, L. B., Vasquez, J. R. & Craik, C. S. (1990) *Proc. Natl. Acad. Sci. USA* 87, 6659–6663.
19. Unal, A., Pray, T. R., Lagunoff, M., Pennington, M. W., Ganem, D. & Craik, C. S. (1997) *J. Virol.* 71, 7030–7038.
20. Jameson, G. W., Roberts, D. V., Adams, R. W., Kyle, W. S. A. & Elmore, D. T. (1973) *Biochem. J.* 131, 107–117.
21. Morrison, J. F. (1969) *Biochim. Biophys. Acta* 185, 269–286.
22. Williams, J. W. & Morrison, J. F. (1979) *Methods Enzymol.* 63, 437–467.
23. Frohman, M. A. (1993) *Methods Enzymol.* 218, 340–356.
24. Kozak, M. (1991) *J. Cell Biol.* 115, 887–903.
25. Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K. & Davie, E. W. (1988) *Biochemistry* 27, 1067–1074.
26. Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W. & Sadler, J. E. (1994) *Proc. Natl. Acad. Sci. USA* 91, 7588–7592.
27. Bork, P. & Beckmann, G. (1993) *J. Mol. Biol.* 231, 539–545.
28. Varela, P. F., Romero, A., Sanz, L., Romao, M. J., Topfer-Petersen, E. & Calvete, J. J. (1997) *J. Mol. Biol.* 274, 635–649.
29. Krieger, M. & Herz, J. (1994) *Annu. Rev. Biochem.* 63, 601–637.
30. Seymour, J. L., Lindquist, R. N., Dennis, M. S., Moffat, B., Yansura, D., Reilly, D., Wessinger, M. E. & Lazarus, R. A. (1994) *Biochemistry* 33, 3949–3958.
31. Nagase, H. (1997) *Biol. Chem.* 378, 151–160.
32. Poloni-Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C. & Antonarkis, S. E. (1997) *Genomics* 44, 309–320.
33. Yamakoka, K., Masuda, K., Ogawa, H., Takagi, K., Umemoto, N. & Yasuoka, S. (1998) *J. Biol. Chem.* 273, 11895–11901.
34. Kessler, E. & Adar, R. (1989) *Eur. J. Biochem.* 186, 115–121.
35. Hulmes, D. J. S., Mould, A. P. & Kessler, E. (1997) *Matrix Biol.* 16, 41–45.
36. Strickl, D. K., Kounnas, M. Z. & Argaves, W. S. (1995) *FASEB J.* 9, 890–898.
37. Moestrup, S. K. (1994) *Biochim. Biophys. Acta* 1197, 197–213.
38. Lu, D., Yuan, X., Zheng, X. & Sadler, J. E. (1997) *J. Biol. Chem.* 272, 31293–31300.
39. Perona, J. J. & Craik, C. S. (1995) *Protein Sci.* 4, 337–360.
40. Perona, J. J. & Craik, C. S. (1997) *J. Biol. Chem.* 272, 29987–29990.
41. Schecter, I. & Berger, A. (1967) *Biochem. Biophys. Res. Commun.* 27, 157–162.
42. LaVallie, E. R., Rehmtulla, A., Racie, L. A., DiBlasio, E. A., Ferenz, C., Grant, K. L., Light, A. & McCoy, J. M. (1993) *J. Biol. Chem.* 268, 23311–23317.
43. Light, A. & Fonseca, P. (1984) *J. Biol. Chem.* 259, 13195–13198.
44. Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., *et al.* (1994) *J. Biol. Chem.* 269, 19976–19982.
45. Harris, J. L., Peterson, E. P., Hudig, D., Thornberry, N. A. & Craik, C. S. (1998) *J. Biol. Chem.* 273, 27364–27373.
46. Nevins, J. R. (1983) *Annu. Rev. Biochem.* 52, 441–466.
47. Kitamoto, Y., Veile, R. A., Donis-Keller, H. & Sadler, J. E. (1995) *Biochemistry* 34, 4562–4568.
48. Beckmann, G. & Bork, P. (1993) *Trends Biochem. Sci.* 18, 40–41.
49. Emi, M., Nakamura, Y., Ogawa, M., Yamamoto, T., Nishide, T., Mori, T. & Matsubara, K. (1986) *Gene* 41, 305–310.
50. Vanderslice, P., Ballinger, S. M., Tam, E. K., Goldstein, S. M., Craik, C. S. & Caghey, G. H. (1990) *Proc. Natl. Acad. Sci. USA* 87, 3811–3815.
51. Tomita, N., Izumoto, Y., Horii, A., Doi, S., Yokouchi, H., Ogawa, M., Mori, T. & Matsubara, K. (1989) *Biochem. Biophys. Res. Commun.* 158, 569–575.
52. Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* 228, 815–822.
53. Wozney, J. M., Rosen, V., Celeste, A. J., Mitscock, L. M., Whitters, M. J., Kriz, R. W., Hewick, R. M. & Wang, E. A. (1988) *Science* 242, 1528–1534.
54. Leytus, S. P., Kurachi, K., Sakariassen, K. S. & Davie, E. W. (1986) *Biochemistry* 25, 4855–4863.

Exhibit 4



US005972616A

United States Patent [19][11] **Patent Number:** **5,972,616****O'Brien et al.**[45] **Date of Patent:** **Oct. 26, 1999**

[54] **TADG-15: AN EXTRACELLULAR SERINE
PROTEASE OVEREXPRESSED IN BREAST
AND OVARIAN CARCINOMAS**

Primary Examiner—Sheela Huff

Attorney, Agent, or Firm—Benjamin Aaron Adler

[75] **Inventors:** Timothy J. O'Brien; Hirotoishi
Tanimoto, both of Little Rock, Ark.

[57] **ABSTRACT**

[73] **Assignee:** The Board of Trustees of the
University of Arkansas, Little Rock,
Ark.

The present invention provides a DNA encoding a TADG-15 protein selected from the group consisting of: (a) isolated DNA which encodes a TADG-15 protein; (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-15 protein; and (c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-15 protein. Also provided is a vector capable of expressing the DNA of the present invention adapted for expression in a recombinant cell and regulatory elements necessary for expression of the DNA in the cell.

[21] **Appl. No.:** 09/027,337

[22] **Filed:** Feb. 20, 1998

[51] **Int. Cl.⁶** **C12Q 1/68**

[52] **U.S. Cl.** **435/6; 435/320.1; 435/69.1;
536/23.1; 536/23.5; 530/350**

[58] **Field of Search** **536/23.1, 23.5;
530/350; 435/320.1, 6, 71.2, 69.1, 41, 71.1**

11 Claims, 17 Drawing Sheets

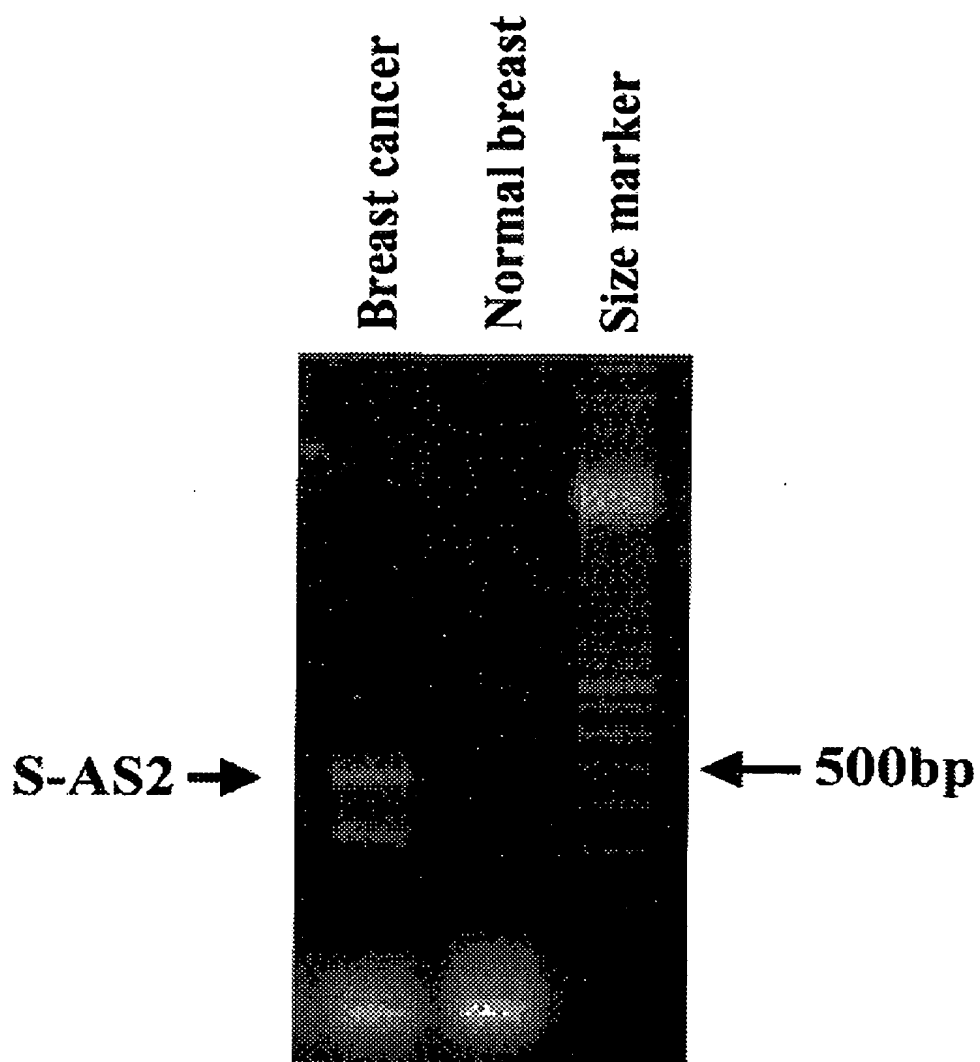


FIG. 1

RIVGGRDTSL GRWPQVSL.RYDG.A HLCGSLISG DWLTAACHF PE....RNRV LSRWRVFAGA VAQASPHGLQ
 RVVGGTDADE GEWPQVSL.HALQG HICGASLISP NWLVSAACHY IDDRGFYSD PTQWTAFLGL HDQSQRSAAPG
 KIIDGAPCAR GSHPWQVAL.LSGNQL H.CGGVLVNE RWLTAACH.K MNEYTVHLGS DTLG..DR.R
 KIVGGYNCEE NSVPYQVSL.NSGYHF ..CGGSLINE QWVVSAGHC.Y KSRIQVRLGE HNIEVLEG.N
 RIVNGEDAVP GSWPWQVSL.QDKTGF HFCGSLISE DWVVTAAHC.GV RTSDVVVAGE FDQGSDEE.N
 RIVGGKVC PK GECPWQVLL.LVNG.A QLCGGTLINT IWVVSAAHCF DKIKNWRNLIAVLGE HDLSEHDGDE
 RIKGGLFADI ASHPWQAAIF AKHRRSPGER FLCGGILISS CWILSAAHCF QERFPPHLL.TVILGR .TYRVVPGE

*

LGVQAVVYHG GYLPERDPNS EENSNDIALV HLSS.PLPLT EYIQPVCLPA ...AGQALVD GKICTVTGWG NTQYYGQQ.A
 VOERRLKRII SHPFENDFTF D...YDIALL ELEK.PAEYS SMVRPICLPD ...ASHVFPA GKAIWVTGWG HTQYGGTG.A
 AQRIKASKSF RHPGYSTQT. ..HVNDMLV KLNS.QARLS SMVKKVRLPS ...RCE..PP GTTCTVSGWG TTTSPDVTFP
 EQFINAAKII RHPQYDRKT. ..LNNDIMLI KLSS.RAVIN ARVSTISLPT ...APP..AT GTKCLISGWG NTASSGADYP
 IQVLKIAKVF KNPKFSILT. ..VNNDITLL KLAT.PARFS QTVSAVCLPS ...ADDDFPA GTLCATTGWG KTKYNANKTP
 QSRRAQVII P....STYVP GTTNHDIALL RLHQ.PVVLTHVWPLCLPE RTFSERTLAF VRFSLVSGWG QLDDRGTAL
 EQKFEVEKYI VHKEFDDDTY D...NDIALL QLKSDSSRCA QESSVVRTVC LPPADLQLPD WTECELSGYG KHEALSPFYS

*

GVLQEARVPI ISNDVCNGAD FYGN..QIKP KMFCAGYPEG G.....IDA CQGDSSGGPFV CEDSISRTPR WRLCGIVSWG
 LILQKGEIRV INQTTCEN LLPQ..QITP RMMCVGFLSG G.....VDS CQGDSSGGPL. ..SSVEADGR IFQAGVVSWG
 SDLMCVDVKL ISPDCTKV. .YKD..LLEN SMLCAGIPDS K.....KNA CNGDSSGGPLV C.....R.... GTLQGLVSWG
 DELQCLDAPV LSQAKCEAS. .YPG..KITS NMFCVGFLEG G.....KDS CQGDSSGGPVV C.....N.... GQLQGVVSWG
 DKLQQAALPL LSNAECKKS. .WGR..RITD VMICAG..AS G.....VSS CMGDSSGGPLV C....QKDG WTLVGIVSWG
 ELMVLNVPR L MTQDCLQQR KVGDSPNITE YMFACAGYSDG S.....KDS CKGDSSGP.. ..HATHYRGT WYLTGIVSWG
 ERLKEAHVRL YPSSRCTSQH LLNRT..VTD NMLCAGDTRS GGQANLHDA CQGDSSGGPLV CLN....DGR MTLVGIISWG

*

T.GCALAQKP	GVYTKVSDFR	EWIFQAIKTH	SEASGMVTQL	~	(SEQ. ID NO: 3)	Heps
D.GCAQRNKP	GVYTRLPLFR	DWIKENTGV~	~	~	(SEQ. ID NO: 14)	Tadg 15
TFPCGQPNDP	GVYTQVCKFT	KWINDTMKKH	R~	~	(SEQ. ID NO: 4)	Scce
D.GCAQKNKP	GVYTKVYNYV	KWIKNTIAAN	S~	~	(SEQ. ID NO: 5)	Try
SDTCS.TSSP	GVYARVTCLI	PWVQKILAN	~	~	(SEQ. ID NO: 6)	Chymb
Q.GCATVGHF	GVYTRVSQYI	EWLQKLMRSE	PRPGVLLRAP	FP	(SEQ. ID NO: 7)	Fac 7
.LGCQKQDVP	GVYTKVTNYL	DWIRDNM RP~	~	~	(SEQ. ID NO: 8)	Tpa

FIG. 2

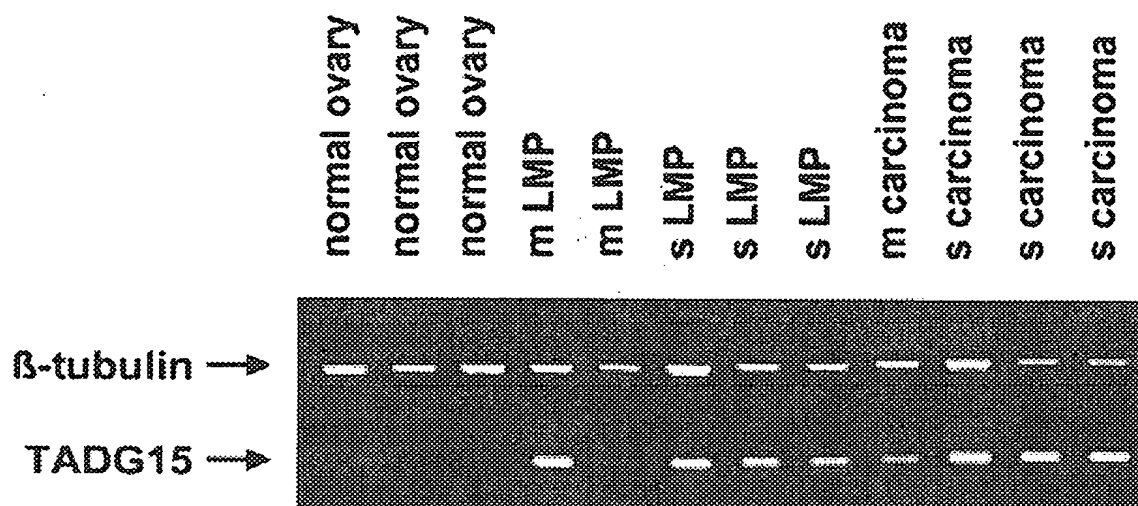


FIG. 3

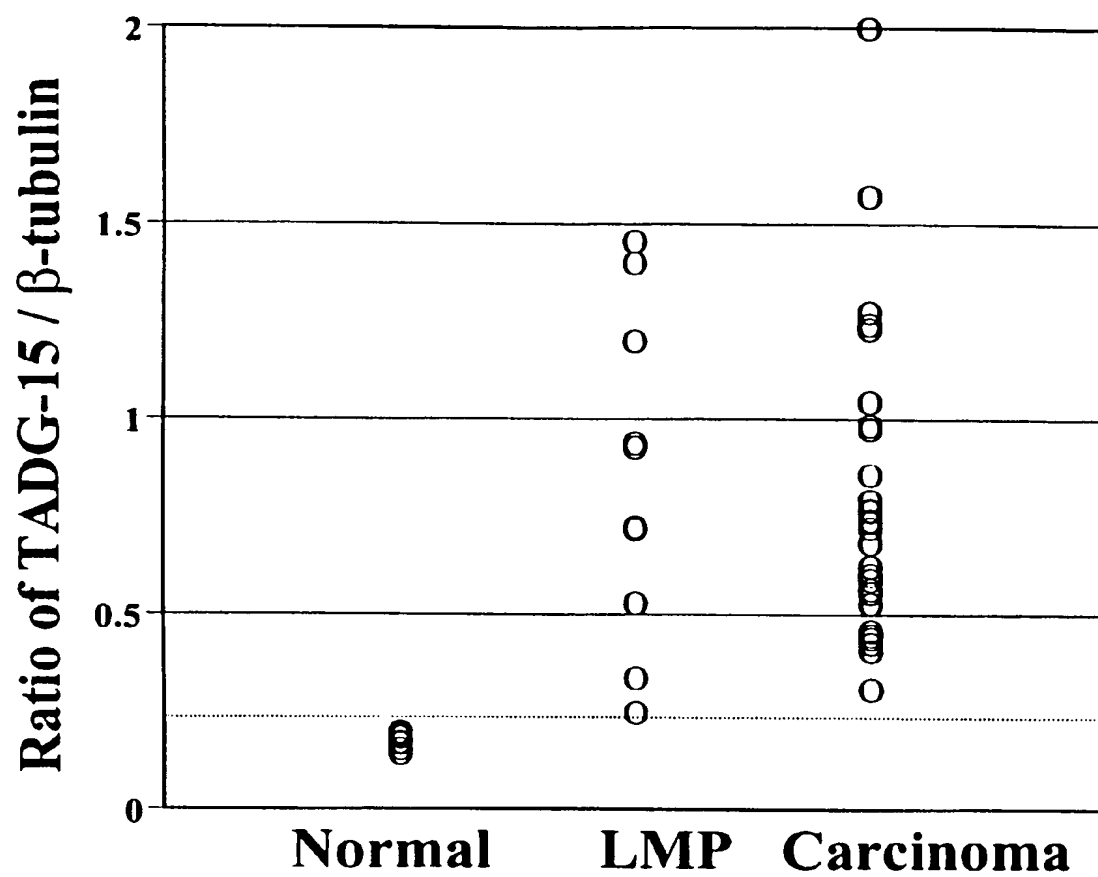


FIG. 4

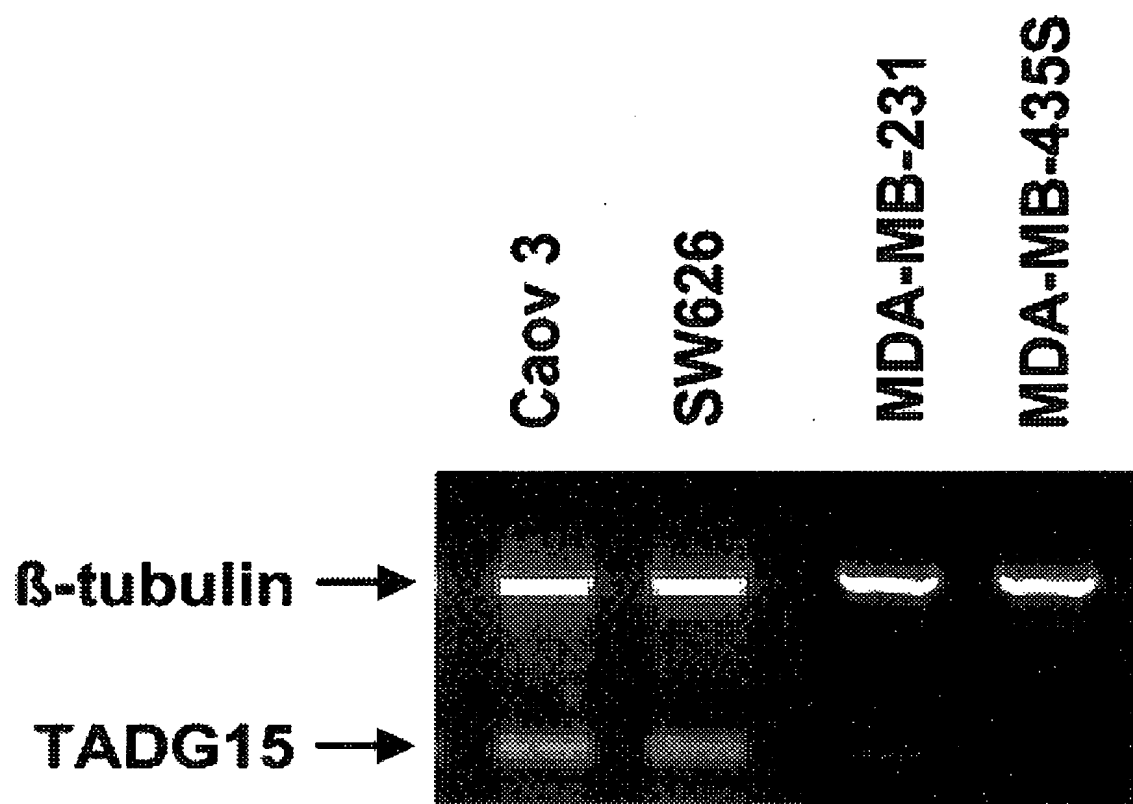


FIG. 5

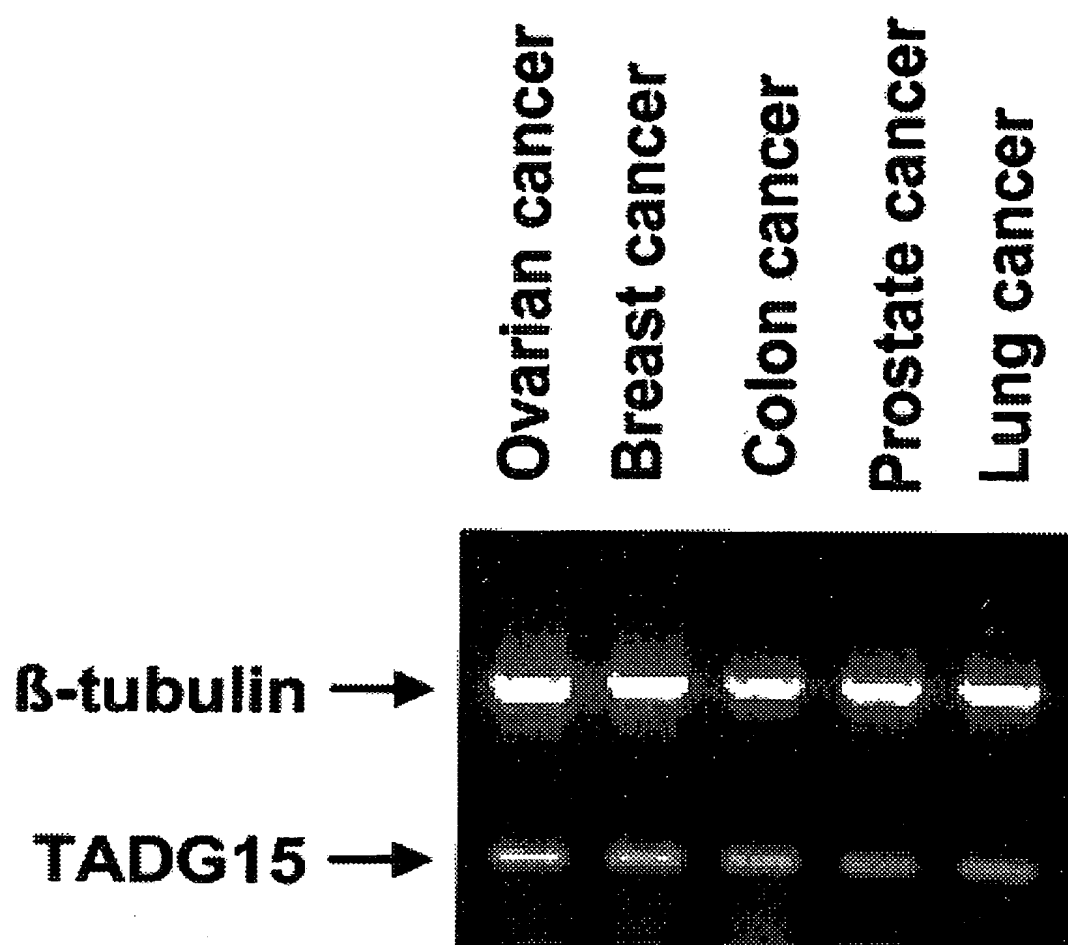


FIG. 6

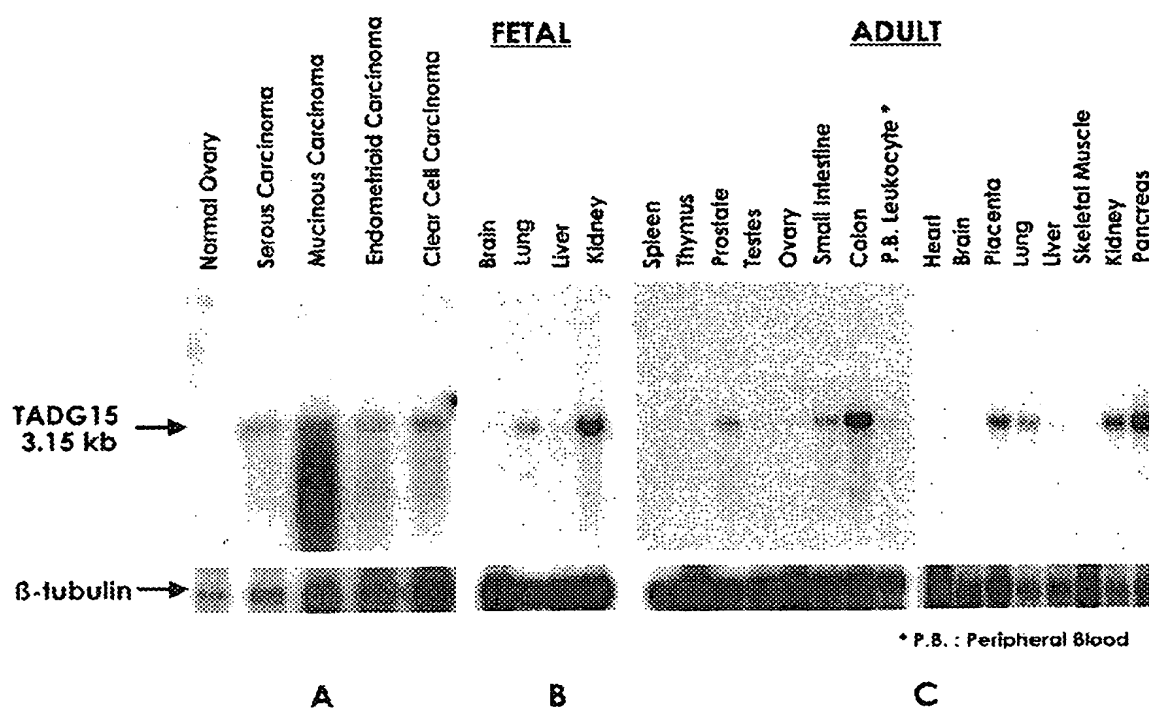


FIG. 7



FIG. 8

1 TCAAGAGCGGCTCGGGGTACCATGGGAGCGATCGGGCCCGCAAGGGCGGAGGGGGCCCGAAGGACTTCGGCGCGGACTC
M G S D R A R K G G G P K D F G A G L
83 AAGTACAACCTCCCGGCACGAGAAAGTGAATGGCTTGGAGGAAGCGGTGGAGTTCCTGCCAGTCAACAACGTCAAGAAGTG
K Y N S R H E K V N G L E E G V E F L P V N N V K K V
164 GAAAAGCATGGCCCGGCGCTGGGTGGTGGCAGCCGTGCTGATCGGCCTCCTCTTGGTCTTGGGATCGGCTTC
E K H G P G R W V V L A A V L I G L L L V L L G I G F
=====

245 CTGGTGTGGCATTTGCAGTACCGGGACGTGCGTGTCCAGAAGTCTTCAATGGCTACATGAGGATCACAAATGAGAAATTT
L V W H L Q Y R D V R V Q K V F N G Y M R I T N E N F
=====

326 GTGGATGCCCTACGAGAACTCCAACCTCACTGAGTTTGTAAAGCCTGGCCAGCAAGGTGAAGGACGGCTGAAGCTGCTGTAC
V D A Y E N S N S T E F V S L A S K V K D A L K L L Y
407 AGCGGAGTCCCATTCCTGGGCCCTTACCACAAGGAGTGGCTGTGACGGCTTCAGCGAGGCGAGCTCATCGCCTACTAC
S G V P F L G P Y H K E S A V T A F S E G S V I A Y Y
488 TGGTCTGAGTTCAGCATCCCGCAGCACCTGGTGGAGGAGCGCGCTCATGGCCGAGGAGCGGTAGTCATGCTGCGCC
W S E F S I P Q H L V E E A E R V M A E E R V V M L P
569 CCGCGGGCGGCTCCCTGAAGTCCCTTTGTGGTCACCTCAGTGGTGGCTTCCCCACGGACTCCAAAACAGTACAGAGGACC
P R A R S L K S F V V T S V V A F P T D S K T V Q R T
650 CAGGACAACAGCTGCAGCTTTGGCCTGCACGCCCGCGGTGTGGAGCTGATGCGCTTCACCACGCCCGGCTTCCCTGACAGC
Q D N S C S F G L H A R G V E L M R F T T P G F P D S
731 CCCTACCCCGCTCATGCCCGCTGCCAGTGGGCCCTGGCGGGGACGCCGACTCAGTGTGAGCCTCACCTTCCGCGAGCTTT
P Y P A H A R C Q W A L R G D A D S V L S L T F R S F
812 GACCTTGGCTCCTGCGACGAGCGGCGAGCACTGGTGACGGTGTACAACACCTGAGCCCATGGAGCCCGACGCCCTG
D L A S C D E R G S D L V T V Y N T L S P M E P H A L
893 GTGCAGTTGTGTGGCACCTACCTCCCTCCTACACCTGACCTTCCACTCCTCCAGAACGTCCTGCTCATCACACTGATA
V Q L C G T Y P P S Y N L T F H S S Q N V L L I T L I

FIG. 9-1

974 ACCAACACTGAGCGGGGCATCCGGGCTTTGAGGCCACCTTCTTCCAGCTGCCTAGGATGAGCAGCTGTGGAGGCCGCTTA
T N T E R R H P G F E A T F F Q L P R M S S C G G R L
1055 CGTAAAGCCCCAGGGACATTCAACAGCCCCCTACTACCCAGGCCACTACCCACCCCAACATTGACTGCACATGGAACATTGAG
R K A Q G T F N S P Y Y P G H Y P P N I D C T W N I E
1136 GTGCCCAACAACCAGCATGTGAAGGTGAGCTTCAAAATTTCTTACCTGTGAGCCCCGGCGTGCCTGCGGGCACCTGCCCC
V P N N Q H V K V S F K F F Y L L E P G V P A G T C P
1217 AAGGACTACGTGGAGATCAATGGGAGAAATACTGCGGAGAGAGTCCCAGTTTCGTCGTCAACCAGCAACAGCAACAAGATC
K D Y V E I N G E K Y C G E R S Q F V V T S N S N K I
1298 ACAGTTCGCTTCCACTCAGATCAGTCCCTACACCGACACCGGCTTCTTAGCTGAATACCTCTCCTACGACTCCAGTGACCCA
T V R F H S D Q S Y T D T G F L A E Y L S Y D S S D P
1379 TGCCCCGGGCGAGTTACGTGCCGACGGGGGGTGTATCCGGAAGGAGCTGGCTGTGATGGCTGGCGCGACTGCACCGAC
C P G Q F T C R T G R C I R K E L R C D G W A D C T D
1460 CACAGCGATGAGCTCAACTGCACTGCGACGCCGCCACCACTTCACTGCAAGAACAAGTTCTGCAAGCCCCCTCTTCTGG
H S D E L N C S C D A G H Q F T C K N K F C K P L F W
1541 GTCTGCGACAGTGTGAACGACTGCGGAGACAACAGCAGCAGGGGTGCAGTTGTCCGGCCCCAGACCTTCAGGTGTTC
V C D S V N D C G D N S D E Q G C S C P A Q T F R C S
1622 AATGGGAAGTGCCTCTCGAAAAGCCAGCAGTGCAATGGGAAGGACGACTGTGGGGACGGGTCCGACGAGGCCCTCCTGCCCC
N G K C L S K S Q Q C N G K D D C G D G S D E A S C P
1703 AAGGTGAACGTCTGTAACCAACACACCTACCGCTGCCCTCAATGGGCTCTGCTTGAGCAAGGGCAACCCCTGAGTGT
K V N V V T C T K H T Y R C L N G L C L S K G N P E C
1784 GACGGGAAGGAGGACTGTAGCGACGGCTCAGATGAGAAGGACTGCGACTGTGGGCTCGGTTCATTACGAGACAGGCTCGT
D G K E D C S D G S D E K D C D C G L R S F T R Q A R
1865 GTTGTGGGGCACGGATGCGGATGAGGGCGAGTGGCCCCCTGGCAGTAAGCCTGCATGCTCTGGGCCAGGGCCACATCTGC
V V G G T D A D E G E W P W Q V S L H A L G Q G H I C
1946 GGTGCTTCCCTCATCTCTCCCAACTGGCTGGTCTCTGCGCGCACACTGTCTACATCGATGACAGAGGATTGAGGTACTCAGAC
G A S L I S P N W L V S A A **H** C Y I D D R G F R Y S D

FIG. 9-2

2027 CCCACGAGTGA CGGCCTTCTGGGCTTGACAGACCAGAGCCAGCGCAGCGCCCTGGGGTGCAGGAGCGCAGGCTCAAG
P T Q W T A F L G L H D Q S Q R S A P G V Q E R R L K
2108 CGCATCATCTCCACCCCTTCTTCAATGACTTCACCTTCGACTATGACATCGCGTCTGCTGGAGCTGGAGAAACCGGCAGAG
R I I S H P F F N D F T F D Y D I A L L E L E K P A E
2189 TACAGCTCCATGGTGGGGCCCATCTGCCTGCCGACGCCCTCCCATGTCTTCCCTGCCGGCAAGGCCATCTGGGTCAACGGGC
Y S S M V R P I C L P D A S H V F P A G K A I W V T G
2270 TGGGACACACCCAGTATGGAGGCACTGGCGGCTGATCCTGCAAAAGGTGAGATCCGGCTCATCAACCAGACCACTGC
W G H T Q Y G G T G A L I L Q K G E I R V I N Q T T C
2351 GAGAACCTCTCTGCCGACAGATCACGCCGCGCATGATGTGCTGGGCTTCCCTCAGCGCGCGTGGACTCCTGCCCAGGGT
E N L L P Q Q I T P R M C V G F L S G G V D S C Q G
2432 GATTCCGGGGACCCCTGTCCAGCGTGGAGCGGATGGCGGATCTTCCAGGCCGCTGTGGTGAGCTGGGGAGACGGCTGC
D S G G P L S S V E A D G R I F Q A G V V S W G D G C
2513 GCTCAGAGGAACAAGCCAGCGGTGTACACAAGGCTCCCTCTGTTCGGGACTGGATCAAAGAGAACTGGGGTATAGGGG
A Q R N K P G V Y T R L P L F R D W I K E N T G V
(SEQ. ID NO: 2)
2594 CCGGGCCACCCAAATGTGTACACCTGCGGGGCCACCCATCGTCCACCCAGTGTGCACGCCCTGCAGGCTGGAGACTGGAC
2675 CGCTGACTGCACAGCGCCCCCAGAACATACACTGTGAATCAATCTCCAGGGCTCCAAATCTGCCCTAGAAAACCTCTCGC
2756 TTCCCTCAGCCTCCAAAGTGGAGCTGGAGGTAGAAGGGAGGACACTGGTGGTTCTACTGACCCAACTGGGGGCCAAAGGTT
2837 TGAAGACACAGCCTCCCCCGCCAGCCCCAAGCTGGGCCGAGGCGGTTTGTGTATATCTGCCTCCCCCTGTCTGTAAAGAGC
2918 AGCGGGAACGGAGCTTCGGAGCCTCCTCAGTGAAGTGGTGGGCTGCCGGATCTGGGCTGTGGGGCCCTTGGGCCACGCT
2999 CTTGAGGAAGCCAGGCTCGGAGGACCCCTGGAAAACAGACGGGTCTGAGACTGAAATTTGTTTACCAGCTCCACAGGTGGA
3080 CTTCAGTGTGTATTGTGTAATGGGTAAACAATTTATTTCTTTTAAAAAATAAAAAAAAAA (SEQ. ID NO: 1)

_____: KOZAK'S CONSENSUS SEQUENCE

====: TRANSMEMBRANE DOMAIN

 : CONSERVED AMINO ACIDS OF CATALYTIC TRIAD H,D,S
FIG. 9-3

1 MGSDRARKGG GGPKDFGAGL KYNSRHEKVN GLEEGVEFLP VNNVKKVEKH 1
 51 GPGHVVVLAA VLIGLLLVLL GIGFLVWHLQ YRDVRVQKVF NGYMRITNEN 2
 101 FVDAYENSNS TEFVSLASKV KDALKLLYSV VPFLGPYHKE SAVTAFSEGS
 151 VIAYYWSEFS IPQHLVEEAE RVMAEERVVM LPPRARSLSK FVVTSVVAFP
 201 TDSKTVQRTQ DNSCSFGLHA RGVELMRFTT PGFPDSPYPA HAR^{*}CQWALRG
 251 DADSVLSLTF RSFDLAS^{*}CDE RGSDLVTVYN TLSPMEPHAL VQL^{*}CGTYPPS 3
 301 Y^{NLT}TFHSSQN VLLITLITNT ERRHPGFEAT FFQLPRMSS^{*}C GGRLRKAQGT
 351 FNSPYYPGHY PPNID^{*}CTWNI EVPNNQHVKV SFKFFYLLEP GVPAGT^{*}CPKD
 401 YVEINGEKY^{*}C GERSQFVVTS NSNKITVRFH SDQSYTDTGF LAEYLSY^{*}DSS
 451 DPCPGQFTCR TGR CIRKELR CDGWADCTDH SDE^{*}LNCSDA GHQFTCKNKF
 501 CKPLFWVCDS VNDCGDN^{SDE} QGCSCPAQTF RCSNGKCLSK SQQCNGKDDC 4
 551 GDG^{SDE}ASCP KVVVVTCTKH TYRCLNGLCL SKGNPECDGK EDCSDG^{SDE}K
 601 DCDGLRSFT RQAR^VVGGTD ADEGEWPQV SLHALGQGHI CGASLISPW
 651 LVSAN^HCYID DRGFYSDPT QWTAFLGLHD QSORSAPGVQ ERRLKRIISH
 701 PFFNDFTFDY ^QIALLELEKP AEYSSMRPI CLPDASHVFP AGKAIWVTGW 5
 751 GHTQYGGTGA LILQKGEIRV INQTTENLL PQQITPRMMC VGFLSGGVDS
 801 CQGD^SGGPLS SVEADGRIFQ AGVVSWDGDC AQRNKPQVYT RLPLFRDWIK
 851 ENTGV (SEQ. ID NO: 2)

* : Conserved cysteine residue

^{NXT} : Possible N-linked glycosylation site

^{SDE} : Conserved SDE motif

^Q : Potential cleavage site

^{H, D, S} : Conserved amino acids of catalytic triad H, D, S

1. Cytoplasmic domain

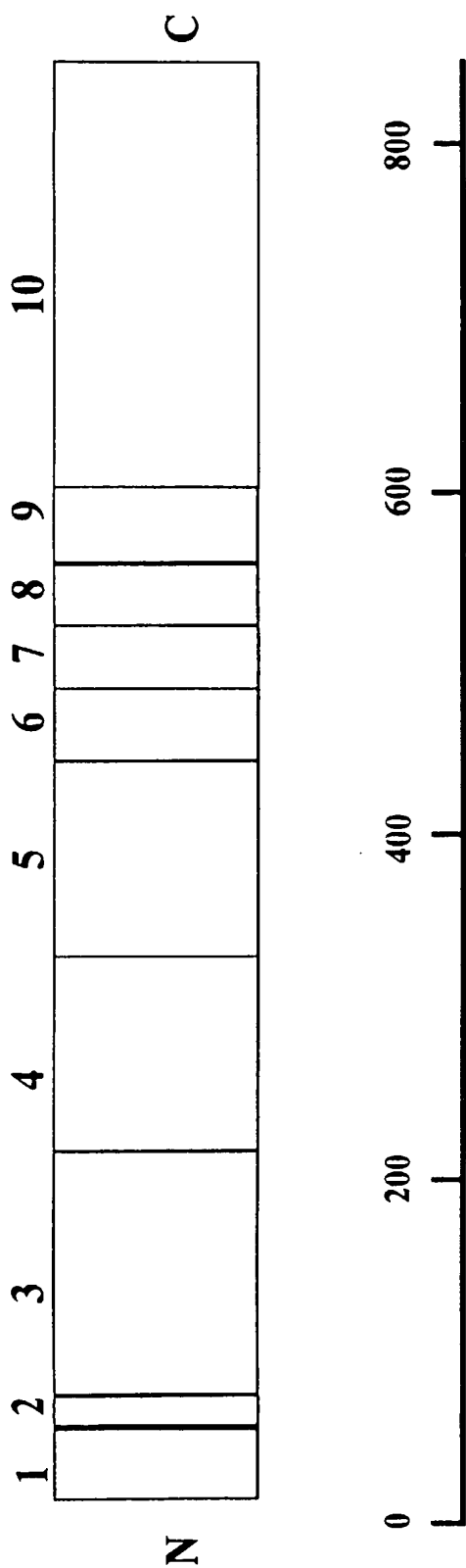
2. Transmembrane domain

3. CUB repeat

4. Ligand-binding repeat (class A motif)
of LDL receptor like domain

5. Serine protease

FIG. 10



1. Cytoplasmic domain
2. Transmembrane domain
3. Extracellular domain
- 4-5. CUB repeat
- 6-9. Ligand-binding repeat (class A motif) of LDL receptor like domain
10. Serine protease

FIG. 11

2374 CACGCCGCCATGATGTGCGTGGGCTTCCTCAGCGGGGGCGTGGACTCCTGCCAGGGTGATTCGGGGGACCCCTGTCCAGCGTGGAGCGGATGGCGG 2473
|||||
2174 CACGCCGCCATGATGTGCGTGGGCTTCCTCAGCGGGGGCGTGGACTCCTGCCAGGGTGATTCGGGGGACCCCTGTCCAGCGTGGAGCGGATGGCGG 2273
|||||
2474 ATCTTCCAGGCCGGTGTGGTGAGCTGGGAGACGGCTGGCGCTCAGAGAACAGCCAGCGGTGTACACAAGGCTCCCTCTGTTTCGGGACTGGATCAAAAG 2573
|||||
2274 ATCTTCCAGGCCGGTGTGGTGAGCTGGGAGAC. GCTGCGCTCAGAGGAACAAGCCAGCGGTGTACACAAGGCTCCCTCTGTTTCGGGAATGGATCAAAAG 2372
|||||
2574 AGAACACTGGGGTATAGGGCCGGGGCCACCCCAAATGTGTACACCTGGGGGGCCACCCATCGTCCACCCAGTGTGCACGCCCTGCAGGCTGGAGACT... 2670
|||||
2373 AGAACACTGGGGTATAGGGCCGGGGCCACCCCAAATGTGTACACCTGGGGGGCCACCCATCGTCCACCCAGTGTGCACGCCCTGCAGGCTGGAGACTCGC 2472
|||||
2671 GGACCGCTGACTGCACCCAGCGCCCCCAGAACATACACTGTGAACCTCAATCTCCAGGGCTCCAAATCTGCTAGAAAACCTCTCGCTTCCTCAGCCTCCAA 2770
|||||
2473 GCACCGTGACCTGCACCCAGCG. CCCCAGAACATACACTGTGAACCTC. ATCTCCAGG..CTCAAACTG. CTAGAAAACCTCTCGCTTCCTCAGCCTCCAA 2567
|||||
2771 AGTGGAGCTGGA. GGTAAGGGGAGG. ACACCTGGTGGTTCTACTGACCCCAACTGGGGGCAAGGTTTGAAGACACAGCCTCCCCCGCCAGCCCCAAGC 2868
|||||
2568 AGTGGAGCTGGAGGGTAGAAGGGGAGGAACACTGGTGGTTCTACTGACCCCAACTGGGG..CAAGGTTTGAAG.CACAG....CTCCGGCAGCCC..AAG 2658
|||||
2869 TGGGCCGAGGCGGGTTTGTGTATATCTGCCCTCCCCTGTCTGTAAAGGAGCAGCGGGAACGGAGCTTCGGAGCCCTCCTCAGTGAAGGTGGTGGGGCTGCCGG 2968
|||||
2659 TGGGCCGAGGACGGGTTTGTGCATA. CTGCC. CTGCTCTATACACGGAAGACCTGGA.....TCTCTAGTGA.....GTGTGACTGCCGG 2735
|||||
2969 ATCTGGGCTGTGGGGCCCTTGGGGCCACGCTCTTGAGGAAGCCCGAGGCTCGGAGGACCCCTGGAAAACAGACGGGTCTGAGACTGAAATTTTACCAGCT 3068
|||||
2736 ATCTGG...CTGTGGTCCCTTGGCCACGCTTCTTGAGGAAGCCCGAGGCTCGGAGGACCCCTGGAAAACAGACGGGTCTGAGACTGAAAATGGTTTACCAGCT 2832
|||||
3069 CCCAGGGTGGACTTCAGTGTGTGATTTTGTGTAATGGGTAAACAATTTATTTCTTTTAAAAAATAAAAAAAAAA 3147 (SEQ. ID NO: 1)
|||||
2833 CCCAGG..TGACTTCAGTGTGTGTA. TTGTGTAATGAGTAAACAATTTATTTCTTTTAAAAAATAAAAAAAAAA..... 2900 (SEQ. ID NO: 9)

FIG. 12-4

TADG-15: AN EXTRACELLULAR SERINE PROTEASE OVEREXPRESSED IN BREAST AND OVARIAN CARCINOMAS

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates generally to the fields of cellular biology and the diagnosis of neoplastic disease. More specifically, the present invention relates to an extracellular serine protease termed Tumor Antigen Derived Gene-15 (TADG-15), which is overexpressed in breast and ovarian carcinomas.

2. Description of the Related Art

Extracellular proteases have been directly associated with tumor growth, shedding of tumor cells and invasion of target organs. Individual classes of proteases are involved in, but not limited to (1) the digestion of stroma surrounding the initial tumor area, (2) the digestion of the cellular adhesion molecules to allow dissociation of tumor cells; and (3) the invasion of the basement membrane for metastatic growth and the activation of both tumor growth factors and angiogenic factors.

The prior art is deficient in the lack of effective means of screening to identify proteases overexpressed in carcinoma. The present invention fulfills this longstanding need and desire in the art.

SUMMARY OF THE INVENTION

The present invention discloses a screening program to identify proteases overexpressed in carcinoma by examining PCR products amplified using differential display in early stage tumors, metastatic tumors compared to that of normal tissues.

In one embodiment of the present invention, there is provided a DNA encoding a TADG-15 protein selected from the group consisting of: (a) isolated DNA which encodes a TADG-15 protein; (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-15 protein; and (c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-15 protein.

In another embodiment of the present invention, there is provided a vector capable of expressing the DNA of the present invention adapted for expression in a recombinant cell and regulatory elements necessary for expression of the DNA in the cell.

In yet another embodiment of the present invention, there is provided a host cell transfected with the vector of the present invention, the vector expressing a TADG-15 protein.

In still yet another embodiment of the present invention, there is provided a method of detecting expression of a TADG-15 mRNA, comprising the steps of: (a) contacting mRNA obtained from the cell with the labeled hybridization probe; and (b) detecting hybridization of the probe with the mRNA.

Other and further aspects, features, and advantages of the present invention will be apparent from the following description of the presently preferred embodiments of the invention given for the purpose of disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

So that the matter in which the above-recited features, advantages and objects of the invention, as well as others

which will become clear, are attained and can be understood in detail, more particular descriptions of the invention briefly summarized above may be had by reference to certain embodiments thereof which are illustrated in the appended drawings. These drawings form a part of the specification. It is to be noted, however, that the appended drawings illustrate preferred embodiments of the invention and therefore are not to be considered limiting in their scope.

FIG. 1 shows a comparison of PCR products derived from normal and breast carcinoma cDNA as shown by staining in an agarose gel.

FIG. 2 shows a comparison of the serine protease catalytic domain of TADG-15 with hepsin (Heps, SEQ ID No: 3), (Scce, SEQ ID No: 4), trypsin (Try, SEQ ID No: 5), chymotrypsin (Chymb, SEQ ID No: 6), factor 7 (Fac7, SEQ ID No: 7) and tissue plasminogen activator (Tpa, SEQ ID No: 8). The asterisks indicate conserved amino acids of catalytic triad.

FIG. 3 shows quantitative PCR analysis of TADG-15 expression.

FIG. 4 shows the ratio of TADG-15 expression to expression of β -tubulin in normal tissues, low malignant potential tumors (LMP) and carcinomas.

FIG. 5 shows the TADG-15 expression in tumor cell lines derived from both ovarian and breast carcinoma tissues.

FIG. 6 shows the overexpression of TADG-15 in other tumor tissues.

FIG. 7 shows the Northern blots of TADG-15 expression in ovarian carcinomas, fetal and normal adult tissues.

FIG. 8 shows a diagram of the TADG-15 transcript and the clones with the origin of their derivation.

FIG. 9 shows nucleotide sequence of the TADG-15 cDNA (SEQ ID No: 1) and amino acid sequence of the TADG-15 protein (SEQ ID No: 2).

FIG. 10 shows the amino acid sequence of the TADG-15 protease including functional sites and domains.

FIG. 11 shows a structure diagram of the TADG-15 protein including functional domains.

FIG. 12 shows a nucleotide sequence comparison between TADG-15 and human SNC-19 (GeneBank accession #U20428).

DETAILED DESCRIPTION OF THE INVENTION

As used herein, the term "cDNA" shall refer to the DNA copy of the mRNA transcript of a gene.

As used herein, the term "derived amino acid sequence" shall mean the amino acid sequence determined by reading the triplet sequence of nucleotide bases in the cDNA.

As used herein the term "screening a library" shall refer to the process of using a labeled probe to check whether, under the appropriate conditions, there is a sequence complementary to the probe present in a particular DNA library. In addition, "screening a library" could be performed by PCR.

As used herein, the term "PCR" refers to the polymerase chain reaction that is the subject of U.S. Pat. Nos. 4,683,195 and 4,683,202 to Mullis, as well as other improvements now known in the art.

The TADG-15 cDNA is 3147 base pairs long (SEQ ID No:1) and encoding for a 855 amino acid protein (SEQ ID No:2). The availability of the TADG-15 gene opens the way for a number studies that can lead to various applications. For example, the TADG-15 gene can be used as a diagnostic

or therapeutic target in ovarian carcinoma and other carcinomas including breast, prostate, lung and colon.

In accordance with the present invention there may be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, e.g., Maniatis, Fritsch & Sambrook, "Molecular Cloning: A Laboratory Manual" (1982); "DNA Cloning: A Practical Approach," Volumes I and II (D. N. Glover ed. 1985); "Oligonucleotide Synthesis" (M. J. Gait ed. 1984); "Nucleic Acid Hybridization" [B. D. Hames & S. J. Higgins eds. (1985)]; "Transcription and Translation" [B. D. Hames & S. J. Higgins eds. (1984)]; "Animal Cell Culture" [R. I. Freshney, ed. (1986)]; "Immobilized Cells And Enzymes" [IRL Press, (1986)]; B. Perbal, "A Practical Guide To Molecular Cloning" (1984).

Therefore, if appearing herein, the following terms shall have the definitions set out below.

The amino acid described herein are preferred to be in the "L" isomeric form. However, residues in the "D" isomeric form can be substituted for any L-amino acid residue, as long as the desired functional property of immunoglobulin-binding is retained by the polypeptide. NH₂ refers to the free amino group present at the amino terminus of a polypeptide. COOH refers to the free carboxy group present at the carboxy terminus of a polypeptide. In keeping with standard polypeptide nomenclature, *J Biol. Chem.*, 243:3552-59 (1969), abbreviations for amino acid residues are shown in the following Table of Correspondence:

TABLE OF CORRESPONDENCE

SYMBOL 1-Letter	3-Letter	AMINO ACID
Y	Tyr	tyrosine
G	Gly	glycine
F	Phe	Phenylalanine
M	Met	methionine
A	Ala	alanine
S	Ser	serine
I	Ile	isoleucine
L	Leu	leucine
T	Thr	threonine
V	Val	valine
P	Pro	proline
K	Lys	lysine
H	His	histidine
Q	Gln	glutamine
E	Glu	glutamic acid
W	Trp	tryptophan
R	Arg	arginine
D	Asp	aspartic acid
N	Asn	asparagine
C	Cys	cysteine

It should be noted that all amino-acid residue sequences are represented herein by formulae whose left and right orientation is in the conventional direction of amino-terminus to carboxy-terminus. Furthermore, it should be noted that a dash at the beginning or end of an amino acid residue sequence indicates a peptide bond to a further sequence of one or more amino-acid residues. The above Table is presented to correlate the three-letter and one-letter notations which may appear alternately herein.

A "replicon" is any genetic element (e.g., plasmid, chromosome, virus) that functions as an autonomous unit of DNA replication in vivo; i.e., capable of replication under its own control.

A "vector" is a replicon, such as plasmid, phage or cosmid, to which another DNA segment may be attached so as to bring about the replication of the attached segment.

A "DNA molecule" refers to the polymeric form of deoxyribonucleotides (adenine, guanine, thymine, or cytosine) in its either single stranded form, or a double-stranded helix. This term refers only to the primary and secondary structure of the molecule, and does not limit it to any particular tertiary forms. Thus, this term includes double-stranded DNA found, inter alia, in linear DNA molecules (e.g., restriction fragments), viruses, plasmids, and chromosomes. In discussing the structure herein according to the normal convention of giving only the sequence in the 5' to 3' direction along the nontranscribed strand of DNA (i.e., the strand having a sequence homologous to the mRNA).

An "origin of replication" refers to those DNA sequences that participate in DNA synthesis.

A DNA "coding sequence" is a double-stranded DNA sequence which is transcribed and translated into a polypeptide in vivo when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5' (amino) terminus and a translation stop codon at the 3' (carboxyl) terminus. A coding sequence can include, but is not limited to, prokaryotic sequences, cDNA from eukaryotic mRNA, genomic DNA sequences from eukaryotic (e.g., mammalian) DNA, and even synthetic DNA sequences. A polyadenylation signal and transcription termination sequence will usually be located 3' to the coding sequence.

Transcriptional and translational control sequences are DNA regulatory sequences, such as promoters, enhancers, polyadenylation signals, terminators, and the like, that provide for the expression of a coding sequence in a host cell.

A "promoter sequence" is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3' direction) coding sequence. For purposes of defining the present invention, the promoter sequence is bounded at its 3' terminus by the transcription initiation site and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence will be found a transcription initiation site, as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase. Eukaryotic promoters often, but not always, contain "TATA" boxes and "CAT" boxes. Prokaryotic promoters contain Shine-Dalgarno sequences in addition to the -10 and -35 consensus sequences.

An "expression control sequence" is a DNA sequence that controls and regulates the transcription and translation of another DNA sequence. A coding sequence is "under the control" of transcriptional and translational control sequences in a cell when RNA polymerase transcribes the coding sequence into mRNA, which is then translated into the protein encoded by the coding sequence.

A "signal sequence" can be included near the coding sequence. This sequence encodes a signal peptide, N-terminal to the polypeptide, that communicates to the host cell to direct the polypeptide to the cell surface or secrete the polypeptide into the media, and this signal peptide is clipped off by the host cell before the protein leaves the cell. Signal sequences can be found associated with a variety of proteins native to prokaryotes and eukaryotes.

The term "oligonucleotide", as used herein in referring to the probe of the present invention, is defined as a molecule comprised of two or more ribonucleotides, preferably more than three. Its exact size will depend upon many factors which, in turn, depend upon the ultimate function and use of the oligonucleotide.

The term "primer" as used herein refers to an oligonucleotide, whether occurring naturally as in a purified restriction digest or produced synthetically, which is capable of acting as a point of initiation of synthesis when placed under conditions in which synthesis of a primer extension product, which is complementary to a nucleic acid strand, is induced, i.e., in the presence of nucleotides and an inducing agent such as a DNA polymerase and at a suitable temperature and pH. The primer may be either single-stranded or double-stranded and must be sufficiently long to prime the synthesis of the desired extension product in the presence of the inducing agent. The exact length of the primer will depend upon many factors, including temperature, source of primer and use the method. For example, for diagnostic applications, depending on the complexity of the target sequence, the oligonucleotide primer typically contains 15-25 or more nucleotides, although it may contain fewer nucleotides.

The primers herein are selected to be "substantially" complementary to different strands of a particular target DNA sequence. This means that the primers must be sufficiently complementary to hybridize with their respective strands. Therefore, the primer sequence need not reflect the exact sequence of the template. For example, a non-complementary nucleotide fragment may be attached to the 5' end of the primer, with the remainder of the primer sequence being complementary to the strand. Alternatively, non-complementary bases or longer sequences can be interspersed into the primer, provided that the primer sequence has sufficient complementarity with the sequence or hybridize therewith and thereby form the template for the synthesis of the extension product.

As used herein, the terms "restriction endonucleases" and "restriction enzymes" refer to enzymes, each of which cut double-stranded DNA at or near a specific nucleotide sequence.

A cell has been "transformed" by exogenous or heterologous DNA when such DNA has been introduced inside the cell. The transforming DNA may or may not be integrated (covalently linked) into the genome of the cell. In prokaryotes, yeast, and mammalian cells for example, the transforming DNA may be maintained on an episomal element such as a plasmid. With respect to eukaryotic cells, a stably transformed cell is one in which the transforming DNA has become integrated into a chromosome so that it is inherited by daughter cells through chromosome replication. This stability is demonstrated by the ability of the eukaryotic cell to establish cell lines or clones comprised of a population of daughter cells containing the transforming DNA. A "clone" is a population of cells derived from a single cell or ancestor by mitosis. A "cell line" is a clone of a primary cell that is capable of stable growth in vitro for many generations.

Two DNA sequences are "substantially homologous" when at least about 75% (preferably at least about 80%, and most preferably at least about 90% or 95%) of the nucleotides match over the defined length of the DNA sequences. Sequences that are substantially homologous can be identified by comparing the sequences using standard software available in sequence data banks, or in a Southern hybridization experiment under, for example, stringent conditions as defined for that particular system. Defining appropriate hybridization conditions is within the skill of the art. See, e.g., Maniatis et al., *supra*; DNA Cloning, Vols. I & II, *supra*; Nucleic Acid Hybridization, *supra*.

A "heterologous" region of the DNA construct is an identifiable segment of DNA within a larger DNA molecule

that is not found in association with the larger molecule in nature. Thus, when the heterologous region encodes a mammalian gene, the gene will usually be flanked by DNA that does not flank the mammalian genomic DNA in the genome of the source organism. In another example, coding sequence is a construct where the coding sequence itself is not found in nature (e.g., a cDNA where the genomic coding sequence contains introns, or synthetic sequences having codons different than the native gene). Allelic variations or naturally-occurring mutational events do not give rise to a heterologous region of DNA as defined herein.

The labels most commonly employed for these studies are radioactive elements, enzymes, chemicals which fluoresce when exposed to ultraviolet light, and others. A number of fluorescent materials are known and can be utilized as labels. These include, for example, fluorescein, rhodamine, auramine, Texas Red, AMCA blue and Lucifer Yellow. A particular detecting material is anti-rabbit antibody prepared in goats and conjugated with fluorescein through an isothiocyanate.

Proteins can also be labeled with a radioactive element or with an enzyme. The radioactive label can be detected by any of the currently available counting procedures. The preferred isotope may be selected from ^3H , ^{14}C , ^{32}P , ^{35}S , ^{36}Cl , ^{51}Cr , ^{57}Co , ^{58}Co , ^{59}Fe , ^{90}Y , ^{125}I , ^{131}I , and ^{186}Re .

Enzyme labels are likewise useful, and can be detected by any of the presently utilized colorimetric, spectrophotometric, fluorospectrophotometric, amperometric or gasometric techniques. The enzyme is conjugated to the selected particle by reaction with bridging molecules such as carbodiimides, diisocyanates, glutaraldehyde and the like. Many enzymes which can be used in these procedures are known and can be utilized. The preferred are peroxidase, β -glucuronidase, β -D-glucosidase, β -D-galactosidase, urease, glucose oxidase plus peroxidase and alkaline phosphatase. U.S. Pat. Nos. 3,654,090, 3,850,752, and 4,016,043 are referred to by way of example for their disclosure of alternate labeling material and methods.

A particular assay system developed and utilized in the art is known as a receptor assay. In a receptor assay, the material to be assayed is appropriately labeled and then certain cellular test colonies are inoculated with a quantity of both the label after which binding studies are conducted to determine the extent to which the labeled material binds to the cell receptors. In this way, differences in affinity between materials can be ascertained.

An assay useful in the art is known as a "cis/trans" assay. Briefly, this assay employs two genetic constructs, one of which is typically a plasmid that continually expresses a particular receptor of interest when transfected into an appropriate cell line, and the second of which is a plasmid that expresses a reporter such as luciferase, under the control of a receptor/ligand complex. Thus, for example, if it is desired to evaluate a compound as a ligand for a particular receptor, one of the plasmids would be a construct that results in expression of the receptor in the chosen cell line, while the second plasmid would possess a promoter linked to the luciferase gene in which the response element to the particular receptor is inserted. If the compound under test is an agonist for the receptor, the ligand will complex with the receptor, and the resulting complex will bind the response element and initiate transcription of the luciferase gene. The resulting chemiluminescence is then measured photometrically, and dose response curves are obtained and compared to those of known ligands. The foregoing protocol is described in detail in U.S. Pat. No. 4,981,784.

As used herein, the term "host" is meant to include not only prokaryotes but also eukaryotes such as yeast, plant and animal cells. A recombinant DNA molecule or gene which encodes a human TADG-15 protein of the present invention can be used to transform a host using any of the techniques commonly known to those of ordinary skill in the art. Especially preferred is the use of a vector containing coding sequences for the gene which encodes a human TADG-15 protein of the present invention for purposes of prokaryote transformation. Prokaryotic hosts may include *E. coli*, *S. typhimurium*, *Serratia marcescens* and *Bacillus subtilis*. Eukaryotic hosts include yeasts such as *Pichia pastoris*, mammalian cells and insect cells.

In general, expression vectors containing promoter sequences which facilitate the efficient transcription of the inserted DNA fragment are used in connection with the host. The expression vector typically contains an origin of replication, promoter(s), terminator(s), as well as specific genes which are capable of providing phenotypic selection in transformed cells. The transformed hosts can be fermented and cultured according to means known in the art to achieve optimal cell growth.

The invention includes a substantially pure DNA encoding a TADG-15 protein, a strand of which DNA will hybridize at high stringency to a probe containing a sequence of at least 15 consecutive nucleotides of (SEQ ID NO: 1). The protein encoded by the DNA of this invention may share at least 80% sequence identity (preferably 85%, more preferably 90%, and most preferably 95%) with the amino acids listed in FIG. 10 (SEQ ID NO:2). More preferably, the DNA includes the coding sequence of the nucleotides of FIG. 9 (SEQ ID NO: 1), or a degenerate variant of such a sequence.

The probe to which the DNA of the invention hybridizes preferably consists of a sequence of at least 20 consecutive nucleotides, more preferably 40 nucleotides, even more preferably 50 nucleotides, and most preferably 100 nucleotides or more (up to 100%) of the coding sequence of the nucleotides listed in FIG. 9 (SEQ ID NO:1) or the complement thereof. Such a probe is useful for detecting expression of TADG-15 in a human cell by a method including the steps of (a) contacting mRNA obtained from the cell with the labeled hybridization probe; and (b) detecting hybridization of the probe with the mRNA.

This invention also includes a substantially pure DNA containing a sequence of at least 15 consecutive nucleotides (preferably 20, more preferably 30, even more preferably 50, and most preferably all) of the region from nucleotides 1 to 3147 of the nucleotides listed in FIG. 9 (SEQ ID NO:1).

By "high stringency" is meant DNA hybridization and wash conditions characterized by high temperature and low salt concentration, e.g., wash conditions of 65° C. at a salt concentration of approximately 0.1×SSC, or the functional equivalent thereof. For example, high stringency conditions may include hybridization at about 42° C. in the presence of about 50% formamide; a first wash at about 65° C. with about 2×SSC containing 1% SDS; followed by a second wash at about 65° C. with about 0.1×SSC.

By "substantially pure DNA" is meant DNA that is not part of a milieu in which the DNA naturally occurs, by virtue of separation (partial or total purification) of some or all of the molecules of that milieu, or by virtue of alteration of sequences that flank the claimed DNA. The term therefore includes, for example, a recombinant DNA which is incorporated into a vector, into an autonomously replicating plasmid or virus, or into the genomic DNA of a prokaryote

or eukaryote; or which exists as a separate molecule (e.g., a cDNA or a genomic or cDNA fragment produced by polymerase chain reaction (PCR) or restriction endonuclease digestion) independent of other sequences. It also includes a recombinant DNA which is part of a hybrid gene encoding additional polypeptide sequence, e.g., a fusion protein. Also included is a recombinant DNA which includes a portion of the nucleotides listed in FIG. 9 (SEQ ID NO: 1) which encodes an alternative splice variant of TADG-15.

The DNA may have at least about 70% sequence identity to the coding sequence of the nucleotides listed in FIG. 9 (SEQ ID NO:1), preferably at least 75% (e.g. at least 80%); and most preferably at least 90%. The identity between two sequences is a direct function of the number of matching or identical positions. When a subunit position in both of the two sequences is occupied by the same monomeric subunit, e.g., if a given position is occupied by an adenine in each of two DNA molecules, then they are identical at that position. For example, if 7 positions in a sequence nucleotides in length are identical to the corresponding positions in a second 10-nucleotide sequence, then the two sequences have 70% sequence identity. The length of comparison sequences will generally be at least 50 nucleotides, preferably at least 60 nucleotides, more preferably at least 75 nucleotides, and most preferably 100 nucleotides. Sequence identity is typically measured using sequence analysis software (e.g., Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, Wis. 53705).

The present invention comprises a vector comprising a DNA sequence which encodes a human TADG-15 protein and said vector is capable of replication in a host which comprises, in operable linkage: a) an origin of replication; b) a promoter; and c) a DNA sequence coding for said protein. Preferably, the vector of the present invention contains a portion of the DNA sequence shown in SEQ ID NO:1. A "vector" may be defined as a replicable nucleic acid construct, e.g., a plasmid or viral nucleic acid. Vectors may be used to amplify and/or express nucleic acid encoding TADG-15 protein. An expression vector is a replicable construct in which a nucleic acid sequence encoding a polypeptide is operably linked to suitable control sequences capable of effecting expression of the polypeptide in a cell. The need for such control sequences will vary depending upon the cell selected and the transformation method chosen.

Generally, control sequences include a transcriptional promoter and/or enhancer, suitable mRNA ribosomal binding sites, and sequences which control the termination of transcription and translation. Methods which are well known to those skilled in the art can be used to construct expression vectors containing appropriate transcriptional and translational control signals. See for example, the techniques described in Sambrook et al., 1989, *Molecular Cloning: A Laboratory Manual* (2nd Ed.), Cold Spring Harbor Press, N.Y. A gene and its transcription control sequences are defined as being "operably linked" if the transcription control sequences effectively control the transcription of the gene. Vectors of the invention include, but are not limited to, plasmid vectors and viral vectors. Preferred viral vectors of the invention are those derived from retroviruses, adenovirus, adeno-associated virus, SV40 virus, or herpes viruses.

By a "substantially pure protein" is meant a protein which has been separated from at least some of those components which naturally accompany it. Typically, the protein is substantially pure when it is at least 60%, by weight, free

from the proteins and other naturally-occurring organic molecules with which it is naturally associated in vivo. Preferably, the purity of the preparation is at least 75%, more preferably at least 90%, and most preferably at least 99%, by weight. A substantially pure TADG-15 protein may be obtained, for example, by extraction from a natural source; by expression of a recombinant nucleic acid encoding an TADG-15 polypeptide; or by chemically synthesizing the protein. Purity can be measured by any appropriate method, e.g., column chromatography such as immunoaffinity chromatography using an antibody specific for TADG-15, polyacrylamide gel electrophoresis, or HPLC analysis. A protein is substantially free of naturally associated components when it is separated from at least some of those contaminants which accompany it in its natural state. Thus, a protein which is chemically synthesized or produced in a cellular system different from the cell from which it naturally originates will be, by definition, substantially free from its naturally associated components. Accordingly, substantially pure proteins include eukaryotic proteins synthesized in *E. coli*, other prokaryotes, or any other organism in which they do not naturally occur.

In addition to substantially full-length proteins, the invention also includes fragments (e.g., antigenic fragments) of the TADG-15 protein (SEQ ID No:2). As used herein, "fragment," as applied to a polypeptide, will ordinarily be at least 10 residues, more typically at least 20 residues, and preferably at least 30 (e.g., 50) residues in length, but less than the entire, intact sequence. Fragments of the TADG-15 protein can be generated by methods known to those skilled in the art, e.g., by enzymatic digestion of naturally occurring or recombinant TADG-15 protein, by recombinant DNA techniques using an expression vector that encodes a defined fragment of TADG-15, or by chemical synthesis. The ability of a candidate fragment to exhibit a characteristic of TADG-15 (e.g., binding to an antibody specific for TADG-15) can be assessed by methods described herein. Purified TADG-15 or antigenic fragments of TADG-15 can be used to generate new antibodies or to test existing antibodies (e.g., as positive controls in a diagnostic assay) by employing standard protocols known to those skilled in the art. Included in this invention are polyclonal antisera generated by using TADG-15 or a fragment of TADG-15 as the immunogen in, e.g., rabbits. Standard protocols for monoclonal and polyclonal antibody production known to those skilled in this art are employed. The monoclonal antibodies generated by this procedure can be screened for the ability to identify recombinant TADG-15 cDNA clones, and to distinguish them from known cDNA clones.

Further included in this invention are TADG-15 proteins which are encoded at least in part by portions of SEQ ID NO:2, e.g., products of alternative mRNA splicing or alternative protein processing events, or in which a section of TADG-15 sequence has been deleted. The fragment, or the intact TADG-15 polypeptide, may be covalently linked to another polypeptide, e.g. which acts as a label, a ligand or a means to increase antigenicity.

The invention also includes a polyclonal or monoclonal antibody which specifically binds to TADG-15. The invention encompasses not only an intact monoclonal antibody, but also an immunologically-active antibody fragment, e.g., a Fab or (Fab)₂ fragment; an engineered single chain Fv molecule; or a chimeric molecule, e.g., an antibody which contains the binding specificity of one antibody, e.g., of murine origin, and the remaining portions of another antibody, e.g., of human origin.

In one embodiment, the antibody, or a fragment thereof, may be linked to a toxin or to a detectable label, e.g. a

radioactive label, non-radioactive isotopic label, fluorescent label, chemiluminescent label, paramagnetic label, enzyme label, or colorimetric label. Examples of suitable toxins include diphtheria toxin, Pseudomonas exotoxin A, ricin, and cholera toxin. Examples of suitable enzyme labels include malate hydrogenase, staphylococcal nuclease, delta-5-steroid isomerase, alcohol dehydrogenase, alpha-glycerol phosphate dehydrogenase, triose phosphate isomerase, peroxidase, alkaline phosphatase, asparaginase, glucose oxidase, beta-galactosidase, ribonuclease, urease, catalase, glucose-6-phosphate dehydrogenase, glucoamylase, acetylcholinesterase, etc. Examples of suitable radioisotopic labels include ³H, ¹²⁵I, ¹³¹I, ³²P, ³⁵S, ¹⁴C, etc.

Paramagnetic isotopes for purposes of in vivo diagnosis can also be used according to the methods of this invention. There are numerous examples of elements that are useful in magnetic resonance imaging. For discussions on in vivo nuclear magnetic resonance imaging, see, for example, Schaefer et al., (1989) *JACC* 14, 472-480; Shreve et al., (1986) *Magn. Reson. Med.* 3, 336-340; Wolf, G. L., (1984) *Physiol. Chem. Phys. Med. NMR* 16, 93-95; Wesbey et al., (1984) *Physiol. Chem. Phys. Med. NMR* 16, 145-155; Runge et al., (1984) *Invest. Radiol.* 19, 408-415. Examples of suitable fluorescent labels include a fluorescein label, an isothiocyanate label, a rhodamine label, a phycoerythrin label, a phycocyanin label, an allophycocyanin label, an ophthaldehyde label, a fluorescamine label, etc. Examples of chemiluminescent labels include a luminal label, an isoluminal label, an aromatic acridinium ester label, an imidazole label, an acridinium salt label, an oxalate ester label, a luciferin label, a luciferase label, an aequorin label, etc.

Those of ordinary skill in the art will know of other suitable labels which may be employed in accordance with the present invention. The binding of these labels to antibodies or fragments thereof can be accomplished using standard techniques commonly known to those of ordinary skill in the art. Typical techniques are described by Kennedy et al., (1976) *Clin. Chim. Acta* 70, 1-31; and Schurs et al., (1977) *Clin. Chim. Acta* 81, 1-40. Coupling techniques mentioned in the latter are the glutaraldehyde method, the periodate method, the dimaleimide method, the m-maleimidobenzyl-N-hydroxy-succinimide ester method. All of these methods are incorporated by reference herein.

Also within the invention is a method of detecting TADG-15 protein in a biological sample, which includes the steps of contacting the sample with the labeled antibody, e.g., radioactively tagged antibody specific for TADG-15, and determining whether the antibody binds to a component of the sample.

As described herein, the invention provides a number of diagnostic advantages and uses. For example, the TADG-15 protein is useful in diagnosing cancer in different tissues since this protein is highly overexpressed in tumor cells. Antibodies (or antigen-binding fragments thereof) which bind to an epitope specific for TADG-15, are useful in a method of detecting TADG-15 protein in a biological sample for diagnosis of cancerous or neoplastic transformation. This method includes the steps of obtaining a biological sample (e.g., cells, blood, plasma, tissue, etc.) from a patient suspected of having cancer, contacting the sample with a labeled antibody (e.g., radioactively tagged antibody) specific for TADG-15, and detecting the TADG-15 protein using standard immunoassay techniques such as an ELISA. Antibody binding to the biological sample indicates that the sample contains a component which specifically binds to an epitope within TADG-15.

Likewise, a standard Northern blot assay can be used to ascertain the relative amounts of TADG-15 mRNA in a cell

or tissue obtained from a patient suspected of having cancer, in accordance with conventional Northern hybridization techniques known to those of ordinary skill in the art. This Northern assay uses a hybridization probe, e.g. radiolabelled TADG-15 cDNA, either containing the full-length, single stranded DNA having a sequence complementary to SEQ ID NO:1 (FIG. 9), or a fragment of that DNA sequence at least 20 (preferably at least 30, more preferably at least 50, and most preferably at least 100 consecutive nucleotides in length). The DNA hybridization probe can be labeled by any of the many different methods known to those skilled in this art.

Antibodies to the TADG-15 protein can be used in an immunoassay to detect increased levels of TADG-15 protein expression in tissues suspected of neoplastic transformation. These same uses can be achieved with Northern blot assays and analyses.

The present invention is directed to DNA encoding a TADG-15 protein selected from the group consisting of: (a) isolated DNA which encodes a TADG-15 protein; (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-15 protein; and (c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-15 protein. Preferably, the DNA has the sequence shown in SEQ ID No:1. More preferably, the DNA encodes a TADG-15 protein having the amino acid sequence shown in SEQ ID No:2.

The present invention is also directed to a vector capable of expressing the DNA of the present invention adapted for expression in a recombinant cell and regulatory elements necessary for expression of the DNA in the cell. Preferably, the vector contains DNA encoding a TADG-15 protein having the amino acid sequence shown in SEQ ID No:2.

The present invention is also directed to a host cell transfected with the vector described herein, said vector expressing a TADG-15 protein. Representative host cells include consisting of bacterial cells, mammalian cells and insect cells.

The present invention is also directed to a isolated and purified TADG-15 protein coded for by DNA selected from the group consisting of: (a) isolated DNA which encodes a TADG-15 protein; (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-15 protein; and (c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-15 protein. Preferably, the isolated and purified TADG-15 protein of claim 9 having the amino acid sequence shown in SEQ ID No:2.

The present invention is also directed to a method of detecting expression of the protein of claim 1, comprising the steps of: (a) contacting mRNA obtained from the cell with the labeled hybridization probe; and (b) detecting hybridization of the probe with the mRNA.

The following examples are given for the purpose of illustrating various embodiments of the invention and are not meant to limit the present invention in any fashion.

EXAMPLE 1

Tissue collection and storage

Upon patient hysterectomy, bilateral salpingo-oophorectomy, or surgical removal of neoplastic tissue, the specimen is retrieved and placed it on ice. The specimen was then taken to the resident pathologist for isolation and identification of specific tissue samples.

Finally, the sample was frozen in liquid nitrogen, logged into the laboratory record and stored at -80°C . Additional specimens were frequently obtained from the Cooperative Human Tissue Network (CHTN). These samples were prepared by the CHTN and shipped to us on dry ice. Upon arrival, these specimens were logged into the laboratory record and stored at -80°C .

EXAMPLE 2

mRNA isolation and cDNA synthesis

Forty-one ovarian tumors (10 low malignant potential tumors and 31 carcinomas) and 10 normal ovaries were obtained from surgical specimens and frozen in liquid nitrogen. The human ovarian carcinoma cell lines SW 626 and Caov 3, the human breast carcinoma cell lines MDA-MB-231 and MDA-MB-435S, and the human uterine cervical carcinoma cell line Hela were purchased from the American Type Culture Collection (Rockville, Md.). Cells were cultured to subconfluency in Dulbecco's modified Eagle's medium, suspended with 10% (v/v) fetal bovine serum and antibiotics.

Messenger RNA (mRNA) isolation was performed according to the manufacturer's instructions using the Mini RiboSep™ Ultra mRNA isolation kit purchased from Becton Dickinson (cat. #30034). This was an oligo(dt) chromatography based system of mRNA isolation. The amount of mRNA recovered was quantitated by UV spectrophotometry.

First strand complementary DNA (cDNA) was synthesized using 5.0 mg of mRNA and either random hexamer or oligo(dT) primers according to the manufacturer's protocol utilizing a first strand synthesis kit obtained from Clontech (cat.# K1402-1). The purity of the cDNA was evaluated by PCR using primers specific for the p53 gene. These primers span an intron such that pure cDNA can be distinguished from cDNA that is contaminated with genomic DNA.

EXAMPLE 3

PCR reactions

The mRNA overexpression of TADG-15 was determined using a quantitative PCR. Oligonucleotide primers were used for: TADG-15, forward 5'-ATGACAGAGGATTTCAGGTAC-3' and reverse 5'-GAAGGTGAAGTCATTGAAGA-3'; and β -tubulin, forward 5'-TGCATTGACAACGAGGC-3' and reverse 5'-CTGTCTTGACATTGTTG-3'. β -tubulin was utilized as an internal control. Reactions were carried out as follows: first strand cDNA generated from 50 ng of mRNA will be used as template in the presence of 1.0 mM MgCl_2 , 0.2 mM dNTPs, 0.025 U Taq polymerase/ml of reaction, and 1xbuffer supplied with enzyme. In addition, primers must be added to the PCR reaction. Degenerate primers which may amplify a variety of cDNAs are used at a final concentration of 2.0 mM each, whereas primers which amplify specific cDNAs are added to a final concentration of 0.2 mM each.

After initial denaturation at 95°C . for 3 minutes, thirty cycles of PCR are carried out in a Perkin Elmer Gene Amp 2400 thermal cycler. Each cycle consists of 30 seconds of denaturation at 95°C ., 30 seconds of primer annealing at the appropriate annealing temperature, and 30 seconds of extension at 72°C . The final cycle will be extended at 72°C . for 7 minutes. To ensure that the reaction succeeded, a fraction of the mixture will be electrophoresed through a 2% agarose/TAE gel stained with ethidium bromide (final concentration 1 mg/ml). The annealing temperature varies according to the primers that are used in the PCR reaction. For the reactions involving degenerate primers, an annealing temperature of 48°C . were used. The appropriate annealing temperature for the TADG-15 and β -tubulin specific primers is 62°C .

13

EXAMPLE 4

T-vector ligation and transformations

The purified PCR products are ligated into the Promega T-vector plasmid and the ligation products are used to transform JM109 competent cells according to the manufacturer's instructions (Promega cat. #A3610). Positive colonies were cultured for amplification, the plasmid DNA isolated by means of the Wizard™ Minipreps DNA purification system (Promega cat #A7500), and the plasmids were digested with *Ap*I and *Sac*I restriction enzymes to determine the size of the insert. Plasmids with inserts of the size(s) visualized by the previously described PCR product gel electrophoresis were sequenced.

EXAMPLE 5

DNA sequencing

Utilizing a plasmid specific primer near the cloning site, sequencing reactions were carried out using PRISM™ Ready Reaction Dye Deoxy™ terminators (Applied Biosystems cat# 401384) according to the manufacturer's instructions. Residual dye terminators were removed from the completed sequencing reaction using a Centri-sep™ spin column (Princeton Separation cat.#CS-901). An Applied Biosystems Model 373A DNA Sequencing System was available and was used for sequence analysis. Based upon the determined sequence, primers that specifically amplify the gene of interest were designed and synthesized.

EXAMPLE 6

Northern blot analysis

10 μ g mRNAs were size separated by electrophoresis through a 1% formaldehyde-agarose gel in 0.02 M MOPS, 0.05 M sodium acetate (pH 7.0), and 0.001 M EDTA. The mRNAs were then blotted to Hybond-N (Amersham) by capillary action in 20 \times SSPE. The RNAs are fixed to the membrane by baking for 2 hours at 80° C.

Additional multiple tissue northern (MTN) blots were purchased from CLONTECH Laboratories, Inc. These blots include the Human MTN blot (cat.#7760-1), the Human MTN II blot (cat.#7759-1), the Human Fetal MTN II blot (cat.#7756-1), and the Human Brain MTN III blot (cat.#7750-1). The appropriate probes were radiolabelled utilizing the Prime-a-Gene Labeling System available from Promega (cat#U1100). The blots were probed and stripped according to the ExpressHyb Hybridization Solution protocol available from CLONTECH (cat.#8015-1 or 8015-2).

EXAMPLE 7

Quantitative PCR

Quantitative-PCR was performed in a reaction mixture consisting of cDNA derived from 50 ng of mRNA, 5 pmol of sense and antisense primers for TADG-15 and the internal control β -tubulin, 0.2 mmol of dNTPs, 0.5 mCi of [α -³²P] dCTP, and 0.625 U of Taq polymerase in 1 \times buffer in a final volume of 25 μ l. This mixture was subjected to 1 minute of denaturation at 95° C. followed by 30 cycles of denaturation for 30 seconds at 95° C., 30 seconds of annealing at 62° C., and 1 minute of extension at 72° C. with an additional 7 minutes of extension on the last cycle. The product was electrophoresed through a 2% agarose gel for separation, the gel was dried under vacuum and autoradiographed. The relative radioactivity of each band was determined by PhosphorImager from Molecular Dynamics.

EXAMPLE 8

The present invention describes the use of primers directed to conserved areas of the serine protease family to

14

identify members of that family which are overexpressed in carcinoma. Several genes were identified and cloned in other tissues, but not previously associated with ovarian carcinoma. The present invention describes a protease identified in ovarian carcinoma. This gene was identified using primers to the conserved area surrounding the catalytic domain of the conserved amino acid histidine and the downstream conserved amino acid serine which lies approximately 150 amino acids towards the carboxyl end of the protease.

The gene encoding the novel extracellular serine protease of the present invention was identified from a group of proteases overexpressed in carcinoma by subcloning and sequencing the appropriate PCR products. An example of such a PCR reaction is given in FIG. 1. Subcloning and sequencing of individual bands from such an amplification provided a basis for identifying the protease of the present invention.

EXAMPLE 9

The sequence determined for the catalytic domain of TADG-15 is presented in FIG. 2 and is consistent with other serine proteases and specifically contains conserved amino acids appropriate for the catalytic domain of the trypsin-like serine protease family. Specific primers (20mers) derived from this sequence were used.

A series of normal and tumor cDNAs were examined to determine the expression of the TADG-15 gene in ovarian carcinoma. In a series of normal derived cDNA compared to carcinoma derived cDNA using β -tubulin as an internal control for PCR amplification, TADG-15 was significantly overexpressed in all of the carcinomas examined and either was not detected or was detected at a very low level in normal epithelial tissue (FIG. 3). This evaluation was extended to a standard panel of about 40 tumors. Using these specific primers, the expression of this gene was also examined in tumor cell lines derived from both ovarian and breast carcinoma tissues as shown in FIG. 5 and in other tumor tissues as shown in FIG. 6. The expression of TADG-15 was also observed in carcinomas of the breast, colon, prostate and lung.

Using the specific sequence for TADG-15 covering the full domain of the catalytic site as a probe for Northern blot analysis, three Northern blots were examined: one derived from ovarian tissues, both normal and carcinoma; one from fetal tissues; and one from adult normal tissues. As shown in FIG. 7, TADG-15 transcripts were noted in all ovarian carcinomas, but were not present in detectable levels in any of the following tissues: a) normal ovary, b) fetal liver and brain, c) adult spleen, thymus, testes, ovary and peripheral blood lymphocytes, d) skeletal muscle, liver, brain or heart. The transcript size was found to be approximately 3.2 kb. The hybridization for the fetal and adult blots was appropriate and done with the same probe as with the ovarian tissue. Subsequent to this examination, it was confirmed that these blots contained other detectable mRNA transcripts.

Initially using the catalytic domain of the protease to probe Hela cDNA and ovarian tumor cDNA libraries, one clone was obtained covering the entire 3' end of the TADG-15 gene from the ovarian tumor library. On further screening using the 5' end of the newly detected clones, two more clones were identified covering the 5' end of the TADG-15 gene from the Hela library (FIG. 8). The complete nucleotide sequence (SEQ ID No: 1) is provided in FIG. 9 along with translation of the open reading frame (SEQ ID No:2).

In the nucleotide sequence, there is a Kozak sequence typical of sequences upstream from the initiation site of

translation. There is also a poly-adenylation signal sequence and a polyadenylated tail. The open reading frame consists of a 855 amino acid sequence (SEQ ID No:2) which includes an amino terminal cytoplasmic tail from amino acids 1-50, an approximately 22 amino acid transmembrane domain followed by an extracellular sequence preceding two CUB repeats identified from complement subcomponents Clr and Cls. These two repeats are followed by four repeat domains of a class A motif of the LDL receptor and these four repeats are followed by the protease enzyme of the trypsin family constituting the carboxyl end of the TADG-15 protein (FIG. 11). Also a clear delineation of the catalytic domain conserved histidine, aspartic acid, serine series along with a series of amino acids conserved in the serine protease family is indicated (FIG. 10).

A search of GeneBank for similar previously identified sequences yielded one such sequence with relatively high homology to a portion of the TADG-15 gene. The similarity between the portion of TADG-15 from nucleotide #182 to 3139 and SNC-19 GeneBank accession #U20428) is approximately 97% (FIG. 12). There are however significant differences between SNC-19 and TADG-15 viz. TADG-15 has an open reading frame of 855 amino acids whereas the longest ORF of SNC-19 is only 173 amino acids. SNC-19 does not include a proper start site for the initiation of translation nor does it include the amino terminal portion of the protein encoded by TADG-15. Moreover, SNC-19 does not include an ORF for a functional serine protease because the His, Asp and Ser residues necessary for function are encoded in different reading frames.

TADG-15 is a highly overexpressed gene in tumors. It is expressed in a limited number of normal tissues, primarily tissues that are involved in either uptake or secretion of molecules e.g. colon and pancreas. TADG-15 is further novel in its component structure of domains in that it has a protease catalytic domain which could be released and used as a diagnostic and which has the potential for a target for therapeutic intervention. TADG-15 also has ligand binding domains which are commonly associated with molecules that internalize or take-up ligands from the external surface of the cell as does the LDL receptor for the LDL cholesterol complex. There is potential that these domains may be involved in uptake of specific ligands and they may offer the potential for making delivery of toxic molecules or genes to tumor cells which express this molecule on their surface. It has features that are similar to the hepsin serine protease molecule in that it also has an amino-terminal transmembrane domain with the proteolytic catalytic domain extended

into the extracellular matrix. The difference here is that TADG-15 includes these ligand binding repeat domains which the hepsin gene does not have. In addition to the use of this gene as a diagnostic or therapeutic target in ovarian carcinoma and other carcinomas including breast, prostate, lung and colon, its ligand-binding domains may be valuable in the uptake of specific molecules into tumor cells. Table 2 shows the number of cases with overexpression of TADG15 in normal ovaries and ovarian tumors.

Any patents or publications mentioned in this specification are indicative of the levels of those skilled in the art to which the invention pertains. These patents and publications are herein incorporated by reference to the same extent as if each individual publication was specifically and individually indicated to be incorporated by reference.

One skilled in the art will readily appreciate that the present invention is well adapted to carry out the objects and obtain the ends and advantages mentioned, as well as those inherent therein. The present examples along with the methods, procedures, treatments, molecules, and specific compounds described herein are presently representative of preferred embodiments, are exemplary, and are not intended as limitations on the scope of the invention. Changes therein and other uses will occur to those skilled in the art which are encompassed within the spirit of the invention as defined by the scope of the claims.

TABLE 2

Number of cases with overexpression of TADG15
in normal ovaries and ovarian tumors.

	N	overexpression of TADG15	expression ratio*
Normal	10	0 (0%)	0.182 ± 0.024
LMP	10	10 (100%)	0.847 ± 0.419
serous	6	6 (100%)	0.862 ± 0.419
mucinous	4	4 (100%)	0.825 ± 0.483
Carcinoma	31	31 (100%)	0.771 ± 0.380
serous	18	18 (100%)	0.779 ± 0.332
mucinous	7	7 (100%)	0.907 ± 0.584
endometrioid	3	3 (100%)	0.502 ± 0.083
clear cell	3	3 (100%)	0.672 ± 0.077

*The ratio of expression level of TADG15 to β -tubulin (mean ± SD)

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 13

<210> SEQ ID NO 1

<211> LENGTH: 3147

<212> TYPE: DNA

<213> ORGANISM: Homo sapiens

<220> FEATURE:

<222> LOCATION: 23..2589

<223> OTHER INFORMATION: cDNA sequence of TADG-15

<400> SEQUENCE: 1

tcaagagcgg cctcggggta ccatggggag cgatcgggcc cgcaaggcg gagggggccc 60
gaaggacttc ggcgaggac tcaagtacaa ctccggcac gagaaagtga atggcttgga 120

-continued

ggaaggcgtg	gagttcctgc	cagtcaacaa	cgtcaagaag	gtggaaaagc	atggcccggg	180
gcgctgggtg	gtgctggcag	ccgtgctgat	cggcctcctc	ttggtcttgc	tggggatcgg	240
cttcctgggtg	tggcattttgc	agtaccggga	cgtgcgtgtc	cagaagggtct	tcaatggcta	300
catgaggatc	acaaatgaga	attttgtgga	tgcctacgag	aactccaact	ccactgagtt	360
tgtaaacctg	gccagcaagg	tgaaggacgc	gctgaagctg	ctgtacagcg	gagtcccatt	420
cctggggccc	taccacaagg	agtcggctgt	gacggccttc	agcgagggca	gcgtcatcgc	480
ctactactgg	tctgagttca	gcaccccgca	gcacctgggtg	gaggaggccg	agcgcgatcat	540
ggccgaggag	cgcgtagtca	tgtctgcccc	gcgggcgcgc	tccctgaagt	cctttgtggt	600
cacctcagtg	gtggctttcc	ccacggactc	caaaacagta	caggaggccc	aggacaacag	660
ctgcagcttt	ggcctgcacg	cccgcggtgt	ggagctgatg	cgcttcacca	cgcgcggctt	720
ccctgacagc	ccctaccccc	ctcatgcccg	ctgccagtg	gccctgcggg	gggacgccga	780
ctcagtgctg	agcctcacct	tccgcagctt	tgccttgctg	tcctgcgacg	agcgcgagcag	840
cgacctgggtg	acggtgtaca	acacctgag	ccccatggag	ccccacgccc	tgggtcagtt	900
gtgtggcacc	taccctccct	cctacaacct	gacctccac	tcctcccaga	acgtcctgct	960
catcacactg	ataaccaaca	ctgagcggcg	gcaccccgcc	tttgaggcca	ccttcttcca	1020
gctgcctagg	atgagcagct	gtggaggccg	cttacgtaaa	gcccagggga	cattcaacag	1080
cccctactac	ccaggccact	acccacccaa	cattgactgc	acatggaaca	ttgaggtgcc	1140
caacaaccag	catgtgaagg	tgagcttcaa	attctctctac	ctgctggagc	ccggcgtgcc	1200
tgcgggcacc	tgccccaaag	actacgtgga	gatcaatggg	gagaaatact	gcggagagag	1260
gtcccagttc	gtcgtcacca	gcaacagcaa	caagatcaca	gttcgcttcc	actcagatca	1320
gtcctacacc	gacaccggct	tcttagctga	atacctctcc	tacgaactcca	gtgacccatg	1380
cccggggcag	ttcacgtgcc	gcacggggcg	gtgtatcccg	aaggagctgc	gctgtgatgg	1440
ctggggccgac	tgcaccgacc	acagcgatga	gctcaactgc	agttgcgacg	ccggccacca	1500
gttcacgtgc	aagaacaagt	tctgcaagcc	cctcttctgg	gtctgcgaca	gtgtgaacga	1560
ctgcggagac	aacagcgacg	agcagggggtg	cagttgtccg	gccagacct	tcagggtgtc	1620
caatgggaag	tgcctctcga	aaagccagca	gtgcaatggg	aaggacgact	gtggggacgg	1680
gtccgacgag	gcctcctgcc	ccaagggtgaa	cgtcgtcact	tgtaccaaac	acacctaccg	1740
ctgcctcaat	gggctctgct	tgagcaagg	caaccctgag	tgtgacggga	aggaggactg	1800
tagcgacggc	tcagatgaga	aggactgcga	ctgtgggctg	cggtcattca	cgagacaggc	1860
tcgtgttgtt	gggggcacgg	atgcggatga	gggcgagtg	ccctggcagg	taagcctgca	1920
tgtctggggc	cagggccaca	tctgcgggtg	ttccctcatc	tctcccaact	ggctgggtctc	1980
tgcgcacac	tgtacatcgc	atgacagagg	attcaggta	tcagacccca	cgcagtggac	2040
ggccttctctg	ggcttgacg	accagagcca	gcgcagcgcc	cctgggggtg	aggagcgag	2100
gctcaagcgc	atcatctccc	accccttctt	caatgacttc	accttcgact	atgacatcgc	2160
gctgctggag	ctggagaaac	cggcagagta	cagctccatg	gtgcggccca	tctgcctgcc	2220
ggacgcctcc	catgtcttcc	ctgccggcaa	ggccatctgg	gtcacgggct	ggggacacac	2280
ccagtatgga	ggcactggcg	cgtctatcct	gcaaaagggt	gagatccgcg	tcataacca	2340
gaccacctgc	gagaacctcc	tgcgcagca	gatcacgcgc	cgcagtatgt	gcgtgggctt	2400
cctcagcggc	ggcgtggact	cctgccagg	tgattccggg	ggaccctgt	ccagcgtgga	2460
ggcgatggg	cggatcttcc	aggccgggt	ggtgagctgg	ggagacggct	gcgctcagag	2520

-continued

```

gaacaagcca ggcgtgtaca caaggctccc tctgtttcgg gactggatca aagagaacac 2580
tggggtatag gggccggggc caccctaatg tgtacacctg cggggccacc catcgtccac 2640
cccagtgtgc acgcctgcag gctggagact ggaccgctga ctgcaccagc gccccagaa 2700
catacactgt gaactcaatc tccagggtc caaatctgcc tagaaaacct ctcgttcct 2760
cagcctccaa agtggagctg ggaggtagaa ggggaggaca ctggtggttc tactgaccca 2820
actgggggca aaggtttgaa gacacagcct cccccgccag cccaagctg ggccgaggcg 2880
cgtttgtgta tatctgcctc ccctgtctgt aaggagcagc gggaacggag cttcgagacc 2940
tcctcagtga aggtggtggg gctgccggat ctgggctgtg gggcccttgg gccacgtct 3000
tgaggaagcc caggctcgga ggaccctgga aaacagacgg gtctgagact gaaattgttt 3060
taccagctcc cagggtggac ttcagtgtgt gtatttgtgt aaatgggtaa aacaatttat 3120
ttctttttta aaaaaaaaaa aaaaaaa 3147

```

```

<210> SEQ ID NO 2
<211> LENGTH: 855
<212> TYPE: PRT
<213> ORGANISM: Homo sapiens
<220> FEATURE:
<223> OTHER INFORMATION: Amino acid sequence of TADG-15 encoded by
nucleotides 23 to 2589 of Sequence 1

```

```

<400> SEQUENCE: 2

```

```

Met Gly Ser Asp Arg Ala Arg Lys Gly Gly Gly Gly Pro Lys Asp
      5              10              15
Phe Gly Ala Gly Leu Lys Tyr Asn Ser Arg His Glu Lys Val Asn
      20              25              30
Gly Leu Glu Glu Gly Val Glu Phe Leu Pro Val Asn Asn Val Lys
      35              40              45
Lys Val Glu Lys His Gly Pro Gly Arg Trp Val Val Leu Ala Ala
      50              55              60
Val Leu Ile Gly Leu Leu Leu Val Leu Leu Gly Ile Gly Phe Leu
      65              70              75
Val Trp His Leu Gln Tyr Arg Asp Val Arg Val Gln Lys Val Phe
      80              85              90
Asn Gly Tyr Met Arg Ile Thr Asn Glu Asn Phe Val Asp Ala Tyr
      95              100             105
Glu Asn Ser Asn Ser Thr Glu Phe Val Ser Leu Ala Ser Lys Val
      110             115             120
Lys Asp Ala Leu Lys Leu Leu Tyr Ser Gly Val Pro Phe Leu Gly
      125             130             135
Pro Tyr His Lys Glu Ser Ala Val Thr Ala Phe Ser Glu Gly Ser
      140             145             150
Val Ile Ala Tyr Tyr Trp Ser Glu Phe Ser Ile Pro Gln His Leu
      155             160             165
Val Glu Glu Ala Glu Arg Val Met Ala Glu Glu Arg Val Val Met
      170             175             180
Leu Pro Pro Arg Ala Arg Ser Leu Lys Ser Phe Val Val Thr Ser
      185             190             195
Val Val Ala Phe Pro Thr Asp Ser Lys Thr Val Gln Arg Thr Gln
      200             205             210
Asp Asn Ser Cys Ser Phe Gly Leu His Ala Arg Gly Val Glu Leu
      215             220             225
Met Arg Phe Thr Thr Pro Gly Phe Pro Asp Ser Pro Tyr Pro Ala
      230             235             240

```

-continued

His Ala Arg Cys Gln Trp Ala Leu Arg Gly Asp Ala Asp Ser Val	245	250	255
Leu Ser Leu Thr Phe Arg Ser Phe Asp Leu Ala Ser Cys Asp Glu	260	265	270
Arg Gly Ser Asp Leu Val Thr Val Tyr Asn Thr Leu Ser Pro Met	275	280	285
Glu Pro His Ala Leu Val Gln Leu Cys Gly Thr Tyr Pro Pro Ser	290	295	300
Tyr Asn Leu Thr Phe His Ser Ser Gln Asn Val Leu Leu Ile Thr	305	310	315
Leu Ile Thr Asn Thr Glu Arg Arg His Pro Gly Phe Glu Ala Thr	320	325	330
Phe Phe Gln Leu Pro Arg Met Ser Ser Cys Gly Gly Arg Leu Arg	335	340	345
Lys Ala Gln Gly Thr Phe Asn Ser Pro Tyr Tyr Pro Gly His Tyr	350	355	360
Pro Pro Asn Ile Asp Cys Thr Trp Asn Ile Glu Val Pro Asn Asn	365	370	375
Gln His Val Lys Val Ser Phe Lys Phe Phe Tyr Leu Leu Glu Pro	380	385	390
Gly Val Pro Ala Gly Thr Cys Pro Lys Asp Tyr Val Glu Ile Asn	395	400	405
Gly Glu Lys Tyr Cys Gly Glu Arg Ser Gln Phe Val Val Thr Ser	410	415	420
Asn Ser Asn Lys Ile Thr Val Arg Phe His Ser Asp Gln Ser Tyr	425	430	435
Thr Asp Thr Gly Phe Leu Ala Glu Tyr Leu Ser Tyr Asp Ser Ser	440	445	450
Asp Pro Cys Pro Gly Gln Phe Thr Cys Arg Thr Gly Arg Cys Ile	455	460	465
Arg Lys Glu Leu Arg Cys Asp Gly Trp Ala Asp Cys Thr Asp His	470	475	480
Ser Asp Glu Leu Asn Cys Ser Cys Asp Ala Gly His Gln Phe Thr	485	490	495
Cys Lys Asn Lys Phe Cys Lys Pro Leu Phe Trp Val Cys Asp Ser	500	505	510
Val Asn Asp Cys Gly Asp Asn Ser Asp Glu Gln Gly Cys Ser Cys	515	520	525
Pro Ala Gln Thr Phe Arg Cys Ser Asn Gly Lys Cys Leu Ser Lys	530	535	540
Ser Gln Gln Cys Asn Gly Lys Asp Asp Cys Gly Asp Gly Ser Asp	545	550	555
Glu Ala Ser Cys Pro Lys Val Asn Val Val Thr Cys Thr Lys His	560	565	570
Thr Tyr Arg Cys Leu Asn Gly Leu Cys Leu Ser Lys Gly Asn Pro	575	580	585
Glu Cys Asp Gly Lys Glu Asp Cys Ser Asp Gly Ser Asp Glu Lys	590	595	600
Asp Cys Asp Cys Gly Leu Arg Ser Phe Thr Arg Gln Ala Arg Val	605	610	615
Val Gly Gly Thr Asp Ala Asp Glu Gly Glu Trp Pro Trp Gln Val	620	625	630
Ser Leu His Ala Leu Gly Gln Gly His Ile Cys Gly Ala Ser Leu			

-continued

125	130	135
Gly Asn Thr Gln Tyr Tyr Gly Gln Gln Ala Gly Val Leu Gln Glu		
140	145	150
Ala Arg Val Pro Ile Ile Ser Asn Asp Val Cys Asn Gly Ala Asp		
155	160	165
Phe Tyr Gly Asn Gln Ile Lys Pro Lys Met Phe Cys Ala Gly Tyr		
170	175	180
Pro Glu Gly Gly Ile Asp Ala Cys Gln Gly Asp Ser Gly Gly Pro		
185	190	195
Phe Val Cys Glu Asp Ser Ile Ser Arg Thr Pro Arg Trp Arg Leu		
200	205	210
Cys Gly Ile Val Ser Trp Gly Thr Gly Cys Ala Leu Ala Gln Lys		
215	220	225
Pro Gly Val Tyr Thr Lys Val Ser Asp Phe Arg Glu Trp Ile Phe		
230	235	240
Gln Ala Ile Lys Thr His Ser Glu Ala Ser Gly Met Val Thr Gln		
245	250	255

Leu

<210> SEQ ID NO 4

<211> LENGTH: 225

<212> TYPE: PRT

<213> ORGANISM: Unknown

<220> FEATURE:

<223> OTHER INFORMATION: Serine protease catalytic domain of Scce homologous to similar domain in TADG-15

<400> SEQUENCE: 4

Lys Ile Ile Asp Gly Ala Pro Cys Ala Arg Gly Ser His Pro Trp		
5	10	15
Gln Val Ala Leu Leu Ser Gly Asn Gln Leu His Cys Gly Gly Val		
20	25	30
Leu Val Asn Glu Arg Trp Val Leu Thr Ala Ala His Cys Lys Met		
35	40	45
Asn Glu Tyr Thr Val His Leu Gly Ser Asp Thr Leu Gly Asp Arg		
50	55	60
Arg Ala Gln Arg Ile Lys Ala Ser Lys Ser Phe Arg His Pro Gly		
65	70	75
Tyr Ser Thr Gln Thr His Val Asn Asp Leu Met Leu Val Lys Leu		
80	85	90
Asn Ser Gln Ala Arg Leu Ser Ser Met Val Lys Lys Val Arg Leu		
95	100	105
Pro Ser Arg Cys Glu Pro Pro Gly Thr Thr Cys Thr Val Ser Gly		
110	115	120
Trp Gly Thr Thr Thr Ser Pro Asp Val Thr Phe Pro Ser Asp Leu		
125	130	135
Met Cys Val Asp Val Lys Leu Ile Ser Pro Gln Asp Cys Thr Lys		
140	145	150
Val Tyr Lys Asp Leu Leu Glu Asn Ser Met Leu Cys Ala Gly Ile		
155	160	165
Pro Asp Ser Lys Lys Asn Ala Cys Asn Gly Asp Ser Gly Gly Pro		
170	175	180
Leu Val Cys Arg Gly Thr Leu Gln Gly Leu Val Ser Trp Gly Thr		
185	190	195
Phe Pro Cys Gly Gln Pro Asn Asp Pro Gly Val Tyr Thr Gln Val		
200	205	210

-continued

Cys Lys Phe Thr Lys Trp Ile Asn Asp Thr Met Lys Lys His Arg
 215 220 225

<210> SEQ ID NO 5
 <211> LENGTH: 225
 <212> TYPE: PRT
 <213> ORGANISM: Unknown
 <220> FEATURE:
 <223> OTHER INFORMATION: Serine protease catalytic domain of trypsin
 (Try) homologous to similar domain in TADG-15

<400> SEQUENCE: 5

Lys Ile Val Gly Gly Tyr Asn Cys Glu Glu Asn Ser Val Pro Tyr
 5 10 15
 Gln Val Ser Leu Asn Ser Gly Tyr His Phe Cys Gly Gly Ser Leu
 20 25 30
 Ile Asn Glu Gln Trp Val Val Ser Ala Gly His Cys Tyr Lys Ser
 35 40 45
 Arg Ile Gln Val Arg Leu Gly Glu His Asn Ile Glu Val Leu Glu
 50 55 60
 Gly Asn Glu Gln Phe Ile Asn Ala Ala Lys Ile Ile Arg His Pro
 65 70 75
 Gln Tyr Asp Arg Lys Thr Leu Asn Asn Asp Ile Met Leu Ile Lys
 80 85 90
 Leu Ser Ser Arg Ala Val Ile Asn Ala Arg Val Ser Thr Ile Ser
 95 100 105
 Leu Pro Thr Ala Pro Pro Ala Thr Gly Thr Lys Cys Leu Ile Ser
 110 115 120
 Gly Trp Gly Asn Thr Ala Ser Ser Gly Ala Asp Tyr Pro Asp Glu
 125 130 135
 Leu Gln Cys Leu Asp Ala Pro Val Leu Ser Gln Ala Lys Cys Glu
 140 145 150
 Ala Ser Tyr Pro Gly Lys Ile Thr Ser Asn Met Phe Cys Val Gly
 155 160 165
 Phe Leu Glu Gly Gly Lys Asp Ser Cys Gln Gly Asp Ser Gly Gly
 170 175 180
 Pro Val Val Cys Asn Gly Gln Leu Gln Gly Val Val Ser Trp Gly
 185 190 195
 Asp Gly Cys Ala Gln Lys Asn Lys Pro Gly Val Tyr Thr Lys Val
 200 205 210
 Tyr Asn Tyr Val Lys Trp Ile Lys Asn Thr Ile Ala Ala Asn Ser
 215 220 225

<210> SEQ ID NO 6
 <211> LENGTH: 231
 <212> TYPE: PRT
 <213> ORGANISM: Unknown
 <220> FEATURE:
 <223> OTHER INFORMATION: Serine protease catalytic domain of
 chymotrypsin (Chymb) homologous to similar domain in TADG-15

<400> SEQUENCE: 6

Arg Ile Val Asn Gly Glu Asp Ala Val Pro Gly Ser Trp Pro Trp
 5 10 15
 Gln Val Ser Leu Gln Asp Lys Thr Gly Phe His Phe Cys Gly Gly
 20 25 30
 Ser Leu Ile Ser Glu Asp Trp Val Val Thr Ala Ala His Cys Gly
 35 40 45

-continued

Val	Arg	Thr	Ser	Asp	Val	Val	Val	Ala	Gly	Glu	Phe	Asp	Gln	Gly	
				50					55					60	
Ser	Asp	Glu	Glu	Asn	Ile	Gln	Val	Leu	Lys	Ile	Ala	Lys	Val	Phe	
				65					70					75	
Lys	Asn	Pro	Lys	Phe	Ser	Ile	Leu	Thr	Val	Asn	Asn	Asp	Ile	Thr	
				80					85					90	
Leu	Leu	Lys	Leu	Ala	Thr	Pro	Ala	Arg	Phe	Ser	Gln	Thr	Val	Ser	
				95					100					105	
Ala	Val	Cys	Leu	Pro	Ser	Ala	Asp	Asp	Asp	Phe	Pro	Ala	Gly	Thr	
				110					115					120	
Leu	Cys	Ala	Thr	Thr	Gly	Trp	Gly	Lys	Thr	Lys	Tyr	Asn	Ala	Asn	
				125					130					135	
Lys	Thr	Pro	Asp	Lys	Leu	Gln	Gln	Ala	Ala	Leu	Pro	Leu	Leu	Ser	
				140					145					150	
Asn	Ala	Glu	Cys	Lys	Lys	Ser	Trp	Gly	Arg	Arg	Ile	Thr	Asp	Val	
				155					160					165	
Met	Ile	Cys	Ala	Gly	Ala	Ser	Gly	Val	Ser	Ser	Cys	Met	Gly	Asp	
				170					175					180	
Ser	Gly	Gly	Pro	Leu	Val	Cys	Gln	Lys	Asp	Gly	Ala	Trp	Thr	Leu	
				185					190					195	
Val	Gly	Ile	Val	Ser	Trp	Gly	Ser	Asp	Thr	Cys	Ser	Thr	Ser	Ser	
				200					205					210	
Pro	Gly	Val	Tyr	Ala	Arg	Val	Thr	Lys	Leu	Ile	Pro	Trp	Val	Gln	
				215					220					225	
Lys	Ile	Leu	Ala	Ala	Asn										
				230											

<210> SEQ ID NO 7

<211> LENGTH: 255

<212> TYPE: PRT

<213> ORGANISM: Unknown

<220> FEATURE:

<223> OTHER INFORMATION: Serine protease catalytic domain of factor 7 (Fac7) homologous to similar domain in TADG-15

<400> SEQUENCE: 7

Arg	Ile	Val	Gly	Gly	Lys	Val	Cys	Pro	Lys	Gly	Glu	Cys	Pro	Trp	
				5					10					15	
Gln	Val	Leu	Leu	Leu	Val	Asn	Gly	Ala	Gln	Leu	Cys	Gly	Gly	Thr	
				20					25					30	
Leu	Ile	Asn	Thr	Ile	Trp	Val	Val	Ser	Ala	Ala	His	Cys	Phe	Asp	
				35					40					45	
Lys	Ile	Lys	Asn	Trp	Arg	Asn	Leu	Ile	Ala	Val	Leu	Gly	Glu	His	
				50					55					60	
Asp	Leu	Ser	Glu	His	Asp	Gly	Asp	Glu	Gln	Ser	Arg	Arg	Val	Ala	
				65					70					75	
Gln	Val	Ile	Ile	Pro	Ser	Thr	Tyr	Val	Pro	Gly	Thr	Thr	Asn	His	
				80					85					90	
Asp	Ile	Ala	Leu	Leu	Arg	Leu	His	Gln	Pro	Val	Val	Leu	Thr	Asp	
				95					100					105	
His	Val	Val	Pro	Leu	Cys	Leu	Pro	Glu	Arg	Thr	Phe	Ser	Glu	Arg	
				110					115					120	
Thr	Leu	Ala	Phe	Val	Arg	Phe	Ser	Leu	Val	Ser	Gly	Trp	Gly	Gln	
				125					130					135	
Leu	Leu	Asp	Arg	Gly	Ala	Thr	Ala	Leu	Glu	Leu	Met	Val	Leu	Asn	
				140					145					150	

-continued

Val	Pro	Arg	Leu	Met	Thr	Gln	Asp	Cys	Leu	Gln	Gln	Ser	Arg	Lys
				155					160					165
Val	Gly	Asp	Ser	Pro	Asn	Ile	Thr	Glu	Tyr	Met	Phe	Cys	Ala	Gly
				170					175					180
Tyr	Ser	Asp	Gly	Ser	Lys	Asp	Ser	Cys	Lys	Gly	Asp	Ser	Gly	Gly
				185					190					195
Pro	His	Ala	Thr	His	Tyr	Arg	Gly	Thr	Trp	Tyr	Leu	Thr	Gly	Ile
				200					205					210
Val	Ser	Trp	Gly	Gln	Gly	Cys	Ala	Thr	Val	Gly	His	Phe	Gly	Val
				215					220					225
Tyr	Thr	Arg	Val	Ser	Gln	Tyr	Ile	Glu	Trp	Leu	Gln	Lys	Leu	Met
				230					235					240
Arg	Ser	Glu	Pro	Arg	Pro	Gly	Val	Leu	Leu	Arg	Ala	Pro	Phe	Pro
				245					250					255

<210> SEQ ID NO 8

<211> LENGTH: 253

<212> TYPE: PRT

<213> ORGANISM: Unknown

<220> FEATURE:

<223> OTHER INFORMATION: Serine protease catalytic domain of tissue plasminogen activator (Tpa) homologous to similar domain in TADG-15

<400> SEQUENCE: 8

Arg	Ile	Lys	Gly	Gly	Leu	Phe	Ala	Asp	Ile	Ala	Ser	His	Pro	Trp
				5					10					15
Gln	Ala	Ala	Ile	Phe	Ala	Lys	His	Arg	Arg	Ser	Pro	Gly	Glu	Arg
				20					25					30
Phe	Leu	Cys	Gly	Gly	Ile	Leu	Ile	Ser	Ser	Cys	Trp	Ile	Leu	Ser
				35					40					45
Ala	Ala	His	Cys	Phe	Gln	Glu	Arg	Phe	Pro	Pro	His	His	Leu	Thr
				50					55					60
Val	Ile	Leu	Gly	Arg	Thr	Tyr	Arg	Val	Val	Pro	Gly	Glu	Glu	Glu
				65					70					75
Gln	Lys	Phe	Glu	Val	Glu	Lys	Tyr	Ile	Val	His	Lys	Glu	Phe	Asp
				80					85					90
Asp	Asp	Thr	Tyr	Asp	Asn	Asp	Ile	Ala	Leu	Leu	Gln	Leu	Lys	Ser
				95					100					105
Asp	Ser	Ser	Arg	Cys	Ala	Gln	Glu	Ser	Ser	Val	Val	Arg	Thr	Val
				110					115					120
Cys	Leu	Pro	Pro	Ala	Asp	Leu	Gln	Leu	Pro	Asp	Trp	Thr	Glu	Cys
				125					130					135
Glu	Leu	Ser	Gly	Tyr	Gly	Lys	His	Glu	Ala	Leu	Ser	Pro	Phe	Tyr
				140					145					150
Ser	Glu	Arg	Leu	Lys	Glu	Ala	His	Val	Arg	Leu	Tyr	Pro	Ser	Ser
				155					160					165
Arg	Cys	Thr	Ser	Gln	His	Leu	Leu	Asn	Arg	Thr	Val	Thr	Asp	Asn
				170					175					180
Met	Leu	Cys	Ala	Gly	Asp	Thr	Arg	Ser	Gly	Gly	Pro	Gln	Ala	Asn
				185					190					195
Leu	His	Asp	Ala	Cys	Gln	Gly	Asp	Ser	Gly	Gly	Pro	Leu	Val	Cys
				200					205					210
Leu	Asn	Asp	Gly	Arg	Met	Thr	Leu	Val	Gly	Ile	Ile	Ser	Trp	Gly
				215					220					225
Leu	Gly	Cys	Gly	Gln	Lys	Asp	Val	Pro	Gly	Val	Tyr	Thr	Lys	Val
				230					235					240

-continued

Thr Asn Tyr Leu Asp Trp Ile Arg Asp Asn Met Arg Pro
 245 250

<210> SEQ ID NO 9
 <211> LENGTH: 2900
 <212> TYPE: DNA
 <213> ORGANISM: Homo sapiens
 <220> FEATURE:
 <223> OTHER INFORMATION: SNC19 mRNA sequence (U20428)

<400> SEQUENCE: 9

```

cgctgggtgg tgctggcagc cgtgctgac gccctcctct tggctcttgc ggggatcggc    60
ttcctggtgt ggcatttgca gtaccgggac gtgcgtgtcc agaaggtctt caatggctac    120
atgaggatca caaatgagaa ttttgtagat gcctacgaga actcctaact cactgagttt    180
gtaagcctgg ccagcaaggt gaaggacgag ctgaagctgc tgtacagcgg agtcccattc    240
ctggggccctt accacaagga gtcggctgtg acggccttca gcgagggcag cgtcatcgcc    300
tactactggt ctgagttcag catcccgag cacctggttg aggaggccga gcgcgtcatg    360
gccaggagcg cgtagtcagt ctgccccgcg gggcgcgctc cctgaagtcc tttgtggtca    420
cctcagtggt ggctttcccc acggactcca aaacagtaca gaggaccag gacaacagct    480
gcagctttgg cctgcacgcc gcggtgtgga gctgatgcgc ttcaccacgc cggcttccct    540
gacagccctt accccgctca tgcccgtgc cagtgggctg cggggacgag acgcagtgc    600
gagctactcg agctgactcg cagcttgact gcgcctcgac gagcgcgaca gcgacctggt    660
gacgtgtaca acacctgag ccccatggag cccacgcct ggtgagtggt tggcacctac    720
ctccctcctt acaacctgac ctccactcc ctcccacgaa cgtcctgctc atcacactga    780
taaccaaacac tgacgcggca tcccggcttt gagggccact tcttccagct gcctaggatg    840
agcagctgtg gagggcgcct acgtaagacc caggggacat tcaacagccc ctactacca    900
ggccactacc caccacaat tgactgcaca tggaaaattg aggtgcccaa caaccagcat    960
gtgaaggtag gcttcaaatt cttctacctg ctggagcccg gcgtgcctgc gggcacctgc   1020
cccaaggact acgtggagat caatggggag aaatactgag gagagaggtc ccagttcgtc   1080
gtcaccagca acagcaacaa gatcacagt cgcttccact cagatcagtc ctacaccgac   1140
accggtctct tagctgaata cctctcctac gactccagt acccatgccc ggggcagttc   1200
acgtgccgca cggggcggtg tatccggaag gagctgcgct gtgatggctg ggcgactgca   1260
ccgaccacag cgatgagctc aactgcagtt gcgacgccg ccaccagttc acgtgcaaga   1320
gcaagtctct caagctcttc tgggtctgag acagtgtgaa cgagtgcgga gacaacagcg   1380
acgagcaggg ttgcatattg ccggaccag accttcaggt gttccaatgg gaagtgcctc   1440
tcgaaaagcc agcagtgcga tgggaaggac gactgtgggg acgggtccga cgaggcctcc   1500
tgccccaagg tgaacgtcgt cacttgtagc aaacacacct accgctgcct caatgggctc   1560
tgcttgagca agggcaaccc tgagtgtgac gggaaggagg actgtagcga cggctcagat   1620
gagaaggact gcgactgtgg gctgcggtca ttcacgagac aggtcgtgtg tgttgggggc   1680
acggatgcgg atgagggcga gtggccctgg caggtaagcc tgcagtctct gggccagggc   1740
cacatctgag gtgcttccct catctctccc aactggctgg tctctgccgc aactgctac   1800
atcgatgaca gaggattcag gtactcagac cccacgcagg acggccttcc tgggcttgca   1860
cgaccagagc cagcgcaggc cctgggggtg aggagcgag gctcaagcgc atcatctccc   1920
accccttctt caatgacttc accttcgact atgacatcgc gctgctggag ctggagaaac   1980

```

-continued

```

cggcagagta cagctccatg gtgcggccca tctgcctgcc ggacgcctgc catgtcttcc 2040
ctgccggcaa ggccatcttg gtcacgggct ggggacacac ccagtatgga ggcaactggc 2100
cgctgatacct gcaaaaggtg gagatccgcg tcatcaacca gaccacctgc gagaacctcc 2160
tgccgcagca gatcacgccg cgcataatgt gcgtgggctt cctcagcggc ggcgtggact 2220
cctgccaggg tgattccggg ggaccctctg ccagcgtgga ggcggatggg cggatcttcc 2280
aggccggtgt ggtgagctgg ggagacgtg cgctcagagg aacaagccag gcgtgtacac 2340
aaggctccct ctgtttcggg aatggatcaa agagaacact ggggtatagg ggcgggggcc 2400
acccaaatgt gtacacctgc gggggcacc ctcgtccacc ccagtgtgca cgcctgcagg 2460
ctggagactc gcgcaccgtg acctgcacca gcgccccaga acatacactg tgaactcatc 2520
tccaggctca aatctgctag aaaacctctc gcttcctcag cctccaaagt ggagctggga 2580
gggtagaagg ggaggaacac tgggtggtct actgacccaa ctggggcaag gtttgaagca 2640
cagctccggc agcccaagtg ggcgaggacg cgtttgtgca tactgccctg ctctatacac 2700
ggaagacctg gatctctagt gagtgtgact gccggatctg gctgtggtcc ttggccacgc 2760
ttcttgagga agccaggctc cggaggaccc tggaaaacag acgggtctga gactgaaaat 2820
ggtttaccag ctccagggtg acttcagtgt gtgtattgtg taaatgagta aaacatttta 2880
tttcttttta aaaaaaaaaa 2900

```

```

<210> SEQ ID NO 10
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<221> NAME/KEY: primer_bind
<222> LOCATION: 1-20
<223> OTHER INFORMATION: Forward primer for analysis of overexpression
of TADG-15 mRNA by quantitative PCR.

```

```

<400> SEQUENCE: 10

```

```

atgacagagg attcaggtac

```

20

```

<210> SEQ ID NO 11
<211> LENGTH: 20
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<221> NAME/KEY: primer_bind
<222> LOCATION: 1-20
<223> OTHER INFORMATION: Reverse primer for analysis of overexpression
of TADG-15 mRNA by quantitative PCR.

```

```

<400> SEQUENCE: 11

```

```

gaaggtgaag tcattgaaga

```

20

```

<210> SEQ ID NO 12
<211> LENGTH: 17
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<221> NAME/KEY: primer_bind
<222> LOCATION: 1-17
<223> OTHER INFORMATION: Forward primer for analysis of B-tubulin mRNA
expression by quantitative PCR.

```

```

<400> SEQUENCE: 12

```

```

tgcatgaca acgaggc

```

17

```

<210> SEQ ID NO 13
<211> LENGTH: 17

```

-continued

<212> TYPE: DNA
 <213> ORGANISM: Artificial Sequence
 <220> FEATURE:
 <221> NAME/KEY: primer_bind
 <222> LOCATION: 1-17
 <223> OTHER INFORMATION: Forward primer for analysis of B-tubulin mRNA
 expression by quantitative PCR.

 <400> SEQUENCE: 13
 ctgtcttgac attgttg

17

What is claimed is:

1. DNA encoding a Tumor Antigen Derived Gene-15 (TADG-15) protein selected from the group consisting of:
 - (a) isolated DNA which encodes a TADG-15 protein;
 - (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-15 protein; and
 - (c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-15 protein.
2. The DNA of claim 1, wherein said DNA has the sequence shown in SEQ ID No:1.
3. The DNA of claim 1, wherein said TADG-15 protein has the amino acid sequence shown in SEQ ID No:2.
4. A vector comprising the DNA of claim 1 and regulatory elements necessary for expression of the DNA in a cell.
5. The vector of claim 4, wherein said DNA encodes a TADG-15 protein having the amino acid sequence shown in SEQ ID No:2.
6. A host cell transfected with the vector of claim 4, said vector expressing a TADG-15 protein.
7. The host cell of claim 6, wherein said cell is selected from group consisting of bacterial cells, mammalian cells, plant cells and insect cells.

8. The host cell of claim 7, wherein said bacterial cell is *E. coli*.
9. Isolated and purified TADG-15 protein coded for by DNA selected from the group consisting of:
 - (a) isolated DNA which encodes a TADG-15 protein;
 - (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-15 protein; and
 - (c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-15 protein.
10. The isolated and purified TADG-15 protein of claim 9 having the amino acid sequence shown in SEQ ID No:2.
11. A method of detecting expression of the protein of claim 9, comprising the steps of:
 - (a) contacting mRNA obtained from a cell with a labeled hybridization probe; and
 - (b) detecting hybridization of the probe with the mRNA.

* * * * *

Exhibit 5

Catalytic mechanism of serine proteases: Reexamination of the pH dependence of the histidyl $^1J_{^{13}\text{C}_2\text{-H}}$ coupling constant in the catalytic triad of α -lytic protease*

(^{13}C NMR/enzyme mechanisms/biosynthetic isotopic enrichment/histidine auxotroph/charge-relay system)

WILLIAM W. BACHOVCHIN[†], ROBERT KAISER[‡], JOHN H. RICHARDS[‡], AND JOHN D. ROBERTS[‡]

[†]Department of Biochemistry and Pharmacology, Tufts University School of Medicine, Boston, Massachusetts 02111; and [‡]The Gates and Crellin Laboratories of Chemistry, California Institute of Technology, Pasadena, California 91125

Contributed by John D. Roberts, August 10, 1981

ABSTRACT L-Histidine, 90% ^{13}C enriched at the C2 position, was incorporated into the catalytic triad of α -lytic protease (EC 3.4.21.12) with the aid of a histidine-requiring mutant of *Lyso bacter enzymogenes* (ATC 29487), and the pH dependence of the coupling constant between this carbon atom and its directly bonded proton was reinvestigated. The high degree of specific ^{13}C isotopic enrichment attainable with the auxotroph permits direct observation and measurement of this coupling constant in proton-coupled ^{13}C NMR spectra at 67.89 MHz and at 15.1 MHz. In contrast to the earlier study, the present results indicate that this coupling constant does respond to a microscopic ionization with pK_a near 7.0; moreover, the magnitude of the values of $^1J_{\text{C-H}}$ observed are in accord with those expected for titration of the histidyl residue. We conclude that the original measurement must be in error and that this coupling constant now also supports a histidyl residue that titrates more or less normally as a component of the catalytic triad of serine proteases.

A "catalytic triad" comprised of the side-chain functional groups of aspartic acid, histidine, and serine has thus far proved to be an invariant feature of the active sites of serine proteinases as demonstrated by x-ray diffraction studies (1-6). The ubiquity and diversity of individual enzymes belonging to this class suggests that this array of Asp-His-Ser residues possesses special catalytic properties. The precise mode of operation of this triad in serine protease-catalyzed hydrolysis of amides and esters is, therefore, of considerable interest.

A prerequisite to the understanding of the effectiveness of this triad is a knowledge of the ionization behavior of its component functional groups, and this has been a controversial issue. A histidyl residue is essential for activity (7-10), and because the activities of serine proteinases increase with pH in a manner indicative of the titration of a single group having a $\text{pK}_a \approx 7.0$ (11), this ionization was originally assumed to represent that of the particular histidyl residue. However, Hunkapiller *et al.* (12) proposed that this pK_a of 7.0 should instead be assigned to the aspartic acid residue and that the histidyl residue should be assigned a pK_a of less than 4.0. The experimental basis for this proposal was a determination that the coupling constant between C2 of the histidyl residue in the catalytic triad of α -lytic protease and its directly bonded proton was independent of pH over the range 4.0-8.0 and indicative of a neutral imidazole ring. The result of this effective reversal of normal pK_a assignments is to make the aspartic acid carboxylate the ultimate charge donor in the operation of the so-called "charge-relay" mechanism (1, 12) of attack on the peptide bond.

The hypothesis that histidyl residues in the catalytic triads of serine proteases are abnormally weak bases, whereas the corresponding aspartic acid residues are abnormally weak acids, has received considerable support, both experimental (13-18) and theoretical (19-23). There are, however, other experimen-

tal results (24-28) that indicate more normal ionization behavior; at one time, substantial controversy on this point existed. Recent ^{15}N (29) and ^1H NMR (30-32) studies strongly indicate that histidyl residues at the catalytic site titrate more or less normally. Nevertheless, the experimental data originally supporting the pK_a -reversal hypothesis remain to be reconciled with these studies. Especially troublesome are the measurements of the histidyl $^1J_{^{13}\text{C}_2\text{-H}}$ coupling constant for α -lytic protease (12) because this result is difficult to attribute to anything but a histidyl residue with an abnormally low pK_a .

The reported measurements of $^1J_{^{13}\text{C}_2\text{-H}}$ are not free of difficulties. A major problem is that the difference in magnitude of this coupling constant between the protonated (≈ 218 Hz) and neutral (≈ 208 Hz) forms of the imidazole ring is small, and its measurement in α -lytic protease was hampered by large line-widths and by background natural-abundance resonances that obscured one line of the doublet. Therefore, determination of the coupling required measurement of $1/2 J$ or the taking of difference spectra. Indeed, whether this measurement could be made with sufficient precision under these circumstances has been questioned (26, 33).

Improved NMR instrumentation operating at higher magnetic field offers the possibility of enhancing the accuracy of the measurements because, at higher fields, interference from background natural-abundance signals should be substantially reduced. Also, a histidine-requiring mutant of *Lyso bacter enzymogenes* is now available which allows one to achieve a higher specific ^{13}C enrichment and, thus, to obtain improved signal detection and resolution. In view of these improved prospects for measuring this coupling constant and the difficulties associated with the earlier study, we report here a reexamination of its pH dependence in α -lytic protease.

MATERIALS AND METHODS

L-Histidine, selectively enriched with ^{13}C at C2 was obtained from Isotope Labelling (Whip, NJ), or KOR Isotopes, (Cambridge, MA), and was synthesized from L-2,5-diamino-4-keto-valeric acid and K^{13}CN as described by Ashley and Harrington (34) and Heath *et al.* (35). Each preparation was judged to be roughly equivalent in regard to purity and specific ^{13}C enrichment ($\approx 92\%$) by ^{13}C NMR spectroscopy. Ac-L-Ala-L-Pro-L-Ala-p-nitroanilide was synthesized as described by Hunkapiller *et al.* (36) and used to assay the activity of the enzyme.

The ^{13}C -labeled histidyl- α -lytic-protease was prepared and purified by culturing a histidine-requiring mutant of *L. enzymogenes* using the previously described procedures (12, 29). The

* Presented in part at the Ninth International Conference on Magnetic Resonance in Biological Systems, Bendor, France, September 1-6, 1980.

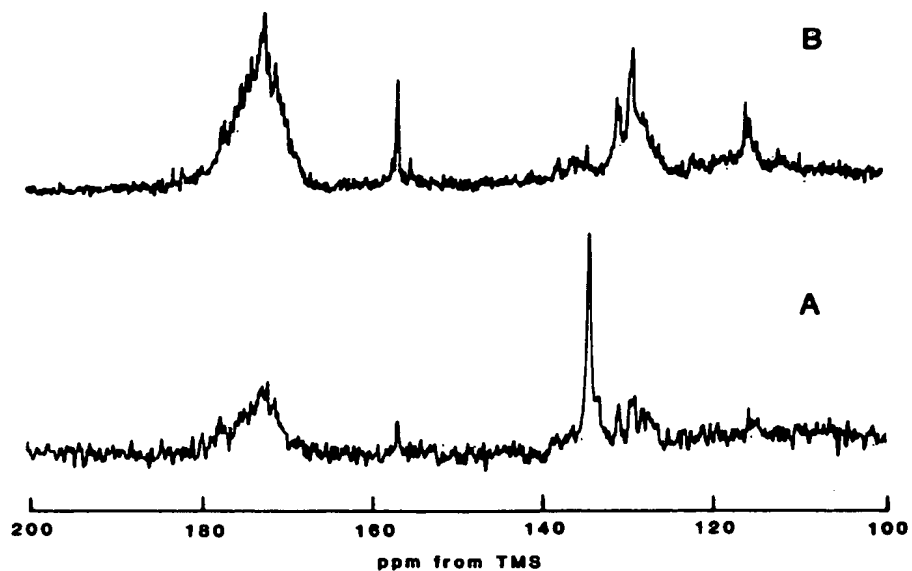


FIG. 1. Proton-decoupled 67.89-MHz ^{13}C NMR spectra of α -lytic protease. (A) $[2\text{-}^{13}\text{C}]\text{Histidyl}$ -enriched α -lytic protease ($\approx 3\text{ mM}$ at pH 4.7; 6400 scans with a recycle time of 0.84 sec). (B) Natural-abundance α -lytic protease ($\approx 8\text{ mM}$ at pH 6.0; 46,000 with a recycle time of 2 sec).

peptidase activity of α -lytic protease was assayed against Ac-L-Ala-L-Pro-L-Ala-*p*-nitroanilide ($4 \times 10^{-4}\text{ M}$ in 0.05 M Tris buffer, pH 8.75, at 25°C). Based on $A_{278}^{1\%} = 8.9$, purified preparations of α -lytic protease used in these NMR studies exhibited k_{cat}/K_m values of $2.0 \times 10^3\text{ M}^{-1}\text{ s}^{-1}$ as compared to a value of

$1.5 \times 10^3\text{ M}^{-1}\text{ s}^{-1}$ reported previously (36).

^{13}C NMR spectra were recorded at 67.89 MHz on a Bruker HX-270 spectrometer and at 15.08 MHz on a Bruker WP-60 spectrometer; 10-mm probes were used with both instruments. The NMR samples were 1–5 mM in α -lytic protease and were

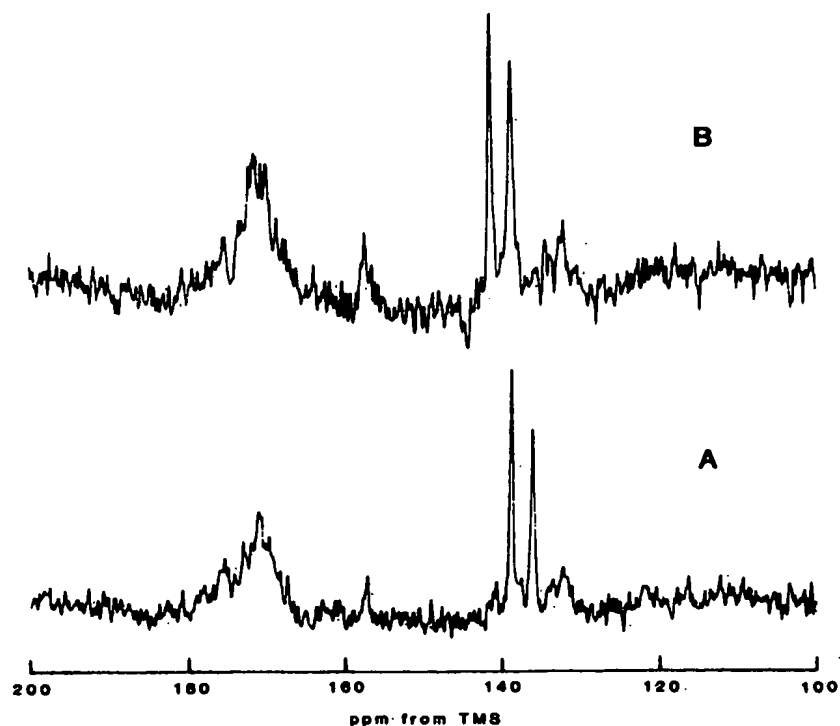


FIG. 2. Proton-coupled 67.89-MHz ^{13}C NMR spectra of $[2\text{-}^{13}\text{C}]\text{Histidyl}$ -enriched α -lytic protease. (A) Enzyme (1.5 mM) at pH 5.54 (25,300 scans with a recycle time of 0.84 sec). (B) Enzyme (1.3 mM) at pH 8.24 (38,500 scans with a recycle time of 0.84 sec).

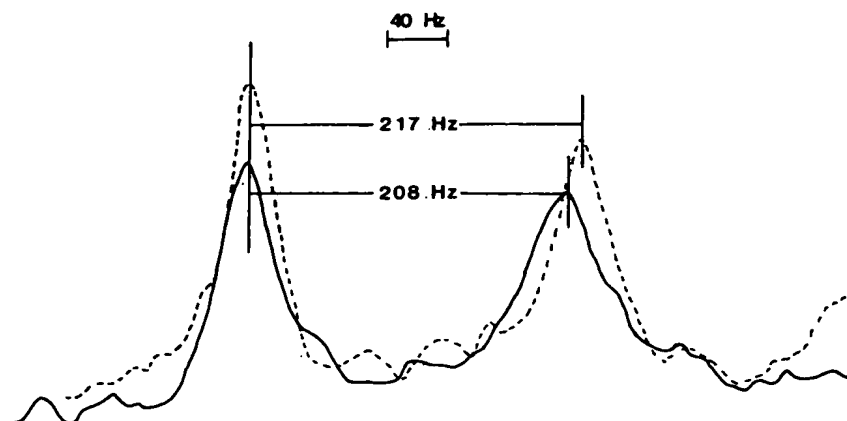


FIG. 3. Comparison of representative high and low pH doublets from 67.89-MHz ^{13}C proton-coupled spectra of $[2-^{13}\text{C}]$ histidyl-enriched α -lytic protease. —, Enzyme (1.34 mM) at pH 8.24 (38,550 scans); ----, 1.5 mM enzyme at pH 5.25 (51,960 scans).

prepared by dissolving lyophilized powders of enzyme in 0.1 M KCl. About 15% of $^2\text{H}_2\text{O}$ was added to provide an internal field frequency lock signal. The relatively sharp signal in ^{13}C NMR spectra of α -lytic protease arising from the guanidinium carbons of the 12 arginine residues (and previously assigned a chemical shift of 157.25 ppm relative to tetramethylsilane) was used as an internal reference after its position relative to internal dioxane was verified to be the same at high and low pH. Chemical shifts are reported in ppm from tetramethylsilane.

In general, 67.89-MHz ^{13}C spectra were acquired by using a 90° radiofrequency pulse (26 μs), a spectral width of 16,000 Hz, and 8000 data points. The ^{13}C spectra at 15.08 MHz were acquired with a 90° pulse (21 μs), a spectral width of 4000 Hz, and 2000 data points.

The pH of the solution and the specific activity of the enzyme were checked both before and after recording each spectrum; only for those samples which exhibited no discernible change in these parameters are spectra reported here. The pH of the sample was varied by the addition of 0.25–0.5 M NaOH or HCl.

RESULTS AND DISCUSSION

Representative proton-decoupled 67.89-MHz ^{13}C NMR spectra of unlabeled α -lytic protease and of $[2-^{13}\text{C}]$ histidyl-labeled α -lytic protease are compared in Fig. 1. The large single resonance at 135 ppm present only in the spectrum of the isotopically enriched enzyme is clearly that of the ^{13}C -labeled carbon of the histidyl residue. The pH dependence of the chemical shift of this resonance is the same as reported earlier (12). Representative proton-coupled ^{13}C NMR spectra at high and low pH are shown in Fig. 2; now both lines of the doublet are clearly resolved at high and low pH, so that $^1J_{\text{C-H}}$ can be measured directly from the peak separation. Six independent determinations of $^1J_{\text{C-H}}$ were made at pH values of 4.66, 5.25, 5.35, 5.47, 5.54, and 6.02, which gave values for $^1J_{\text{C-H}}$ of 219, 217, 219, 217, 217, and 216 Hz, respectively. Two determinations of $^1J_{\text{C-H}}$ at pH 8.24 and 8.44 gave values of 208 and 204, respectively. Either Lorentzian or parabolic interpolation of the peak positions yielded the same value for $^1J_{\text{C-H}}$. The curves in Fig. 3 for representative high and low pH doublets demonstrate that $^1J_{\text{C-H}}$ does change with pH.

In addition to the high-field ^{13}C NMR measurements at 67.89 MHz, the coupling constant was also determined by ^{13}C NMR spectroscopy at 15.1 MHz, and even at this lower magnetic field, both lines of the doublet were sufficiently resolved to

allow direct measurement of the coupling. Two independent determinations of the coupling constant in both the high and low pH ranges gave effectively the same results as the measurements at 67.89 MHz.

The present results indicate that this coupling constant does respond to an ionization of the histidyl residue with a pK_a near 7.0, and the original measurements (12) must be in error. The source of this error is, at present, not clear, but possibly derives from the presence of multiple forms of the enzyme (31) at acidic pH. These forms can be resolved at 125 MHz where they are in slow exchange (R. J. Kaiser and T. G. Perkins, personal communication).

Consequently, the NMR data (^{15}N , ^{13}C , and ^1H) now support a histidyl residue which titrates more or less normally as a component of the active-site catalytic triads of serine proteases—at least for the free enzyme in solution. Other experimental or theoretical studies that support, as well as mechanistic schemes based upon, the pK_a -reversal hypothesis need reappraisal.

This work was supported by grants from the National Institutes of Health (GM-27927 and GM164221) and from Research Corporation. The high-field NMR experiments were performed at the NMR Facility for Biomolecular Research located at the F. Bitter National Magnet Laboratory (Massachusetts Institute of Technology). The NMR Facility is supported by Grant RR00995 from the Division of Research Resources of the National Institutes of Health and by National Science Foundation Contract C-670.

1. Blow, D. M., Birktoft, J. J., & Hartley, B. S. (1969) *Nature (London)* 221, 337–340.
2. Stroud, R. M., Kay, L. M., & Dickerson, R. E. (1974) *J. Mol. Biol.* 83, 185–208.
3. Sawyer, L., Shotton, D. M., Campbell, J. W., Wendell, P. L., Muirhead, H., Watson, H. C., Diamond, R., & Ladner, R. C. (1978) *J. Mol. Biol.* 118, 137–208.
4. Matthews, D. A., Alden, R. A., Birktoft, J. J., Freer, S. T., & Kraut, J. (1977) *J. Biol. Chem.* 252, 8875–8883.
5. Coddling, P. W., Delbaere, L. T. J., Hayakawa, K., Hutcheon, W. L. B., James, M. N. G., & Jurásek, L. (1974) *Can. J. Biochem.* 52, 208–220.
6. James, M. N. G., Delbaere, L. T. J., & Brayer, G. D. (1978) *Can. J. Biochem.* 56, 396–402.
7. Ong, E. B., Shaw, E., & Schoellman, G. (1964) *J. Am. Chem. Soc.* 86, 1271–1272.
8. Schoellman, G., & Shaw, E. (1962) *Biochem. Biophys. Res. Commun.* 7, 36–40.
9. Ray, W. J., Jr. & Koshland, D. E., Jr. (1960) *Brookhaven Symp. Biol.* 13, 135–150.

10. Weil, L., James, S. & Buchert, A. R. (1953) *Arch. Biochem. Biophys.* 46, 266-278.
11. Hess, G. P. (1971) *Enzymes* 3, 213-248.
12. Hunkapiller, M. W., Smallcombe, S. H., Whitaker, D. R. & Richards, J. H. (1973) *Biochemistry* 12, 4732-4743.
13. Koeppe, R. E., II & Stroud, R. M. (1976) *Biochemistry* 15, 3450-3458.
14. Markley, J. L. (1975) *Acc. Chem. Res.* 8, 70-80.
15. Markley, J. L. & Porubcan, M. A. (1976) *J. Mol. Biol.* 102, 487-509.
16. Faraggi, M., Klapper, M. H. & Dorfman, L. M. (1978) *J. Phys. Chem.* 82, 508-512.
17. Komiyama, M., Bender, M. L., Utsaka, M. & Takeda, A. (1977) *Proc. Natl. Acad. Sci. USA* 74, 2634-2638.
18. Komiyama, M., Rosel, T. R. & Bender, M. L. (1977) *Proc. Natl. Acad. Sci. USA* 74, 23-25.
19. Scheiner, S., Kleier, D. A. & Lipscomb, W. N. (1975) *Proc. Natl. Acad. Sci. USA* 72, 2606-2610.
20. Scheiner, S. & Lipscomb, W. N. (1976) *Proc. Natl. Acad. Sci. USA* 73, 432-436.
21. Beppeu, Y. & Yomosa, S. (1977) *J. Phys. Soc. Jpn.* 42, 1694-1700.
22. Amidon, G. L. (1974) *J. Theor. Biol.* 46, 101-109.
23. Kitayama, H. P. & Fukutome, H. (1976) *J. Theor. Biol.* 60, 1-18.
24. Robillard, G. & Shulman, R. G. (1972) *J. Mol. Biol.* 71, 507-511.
25. Robillard, G. & Shulman, R. G. (1974) *J. Mol. Biol.* 86, 519-540.
26. Robillard, G. & Shulman, R. G. (1974) *J. Mol. Biol.* 86, 541-558.
27. Bruice, T. C. (1976) *Annu. Rev. Biochem.* 45, 331-373.
28. Rogers, G. A. & Bruice, T. C. (1974) *J. Am. Chem. Soc.* 96, 2473-2481.
29. Bachovchin, W. W. & Roberts, J. D. (1978) *J. Am. Chem. Soc.* 100, 8041-8047.
30. Markley, J. L. & Ibañez, I. B. (1978) *Biochemistry* 17, 4627-4640.
31. Westler, W. M. (1980) Dissertation (Purdue Univ., Lafayette, IN).
32. Markley, J. L., Neves, D. E., Westler, W. M., Ibañez, I. B., Porubcan, M. A. & Baillargeon, M. W. (1980) *Dev. Biochem.* 10, 31-62.
33. Egan, W., Shindo, H. & Cohen, J. (1977) *Annu. Rev. Biophys. Bioeng.* 6, 383-417.
34. Ashley, J. H. & Harrington, R. (1930) *J. Chem. Soc.*, 2586-2590.
35. Heath, H., Lawson, A. & Rimington, C. (1951) *J. Chem. Soc.*, 2215-2222.
36. Hunkapiller, M. W., Forgac, M. D. & Richards, J. H. (1976) *Biochemistry*, 15, 5581-5588.

Exhibit 6

Perspectives in Bioconjugate Chemistry

EDITED BY
Claude F. Meares
University of California



American Chemical Society, Washington, DC 1993



Library of Congress Cataloging-in-Publication Data

Perspectives in bioconjugate chemistry / edited by Claude F. Meares.

p. cm.

Contains a collection of articles previously published in the journal:
Bioconjugate chemistry.

Includes bibliographical references and index.


ISBN 0-8412-2672-5

1. Bioconjugates.

I. Meares, Claude F., 1946- . II. American Chemical Society.

QP517.B49P47 1993
574.19'2—dc20

93-15385
CIP

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984. 

Copyright © 1993

American Chemical Society

All Rights Reserved. The appearance of the code at the bottom of the first page of each chapter in this volume indicates the copyright owner's consent that reprographic copies of the chapter may be made for personal or internal use or for the personal or internal use of specific clients. This consent is given on the condition, however, that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc., 27 Congress Street, Salem, MA 01970; for copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to copying or transmission by any means—graphic or electronic—for any other purpose, such as for general distribution, for advertising or promotional purposes, for creating a new collective work, for resale, or for information storage and retrieval systems. The copying fee for each chapter is indicated in the code at the bottom of the first page of the chapter.

The citation of trade names and/or names of manufacturers in this publication is not to be construed as an endorsement or as approval by ACS of the commercial products or services referenced herein; nor should the mere reference herein to any drawing, specification, chemical process, or other data be regarded as a license or as a conveyance of any right or permission to the holder, reader, or any other person or corporation, to manufacture, reproduce, use, or sell any patented invention or copyrighted work that may in any way be related thereto. Registered names, trademarks, etc., used in this publication, even without specific indication thereof, are not to be considered unprotected by law.

PRINTED IN THE UNITED STATES OF AMERICA

A Brief Survey of Methods for Preparing Protein Conjugates with Dyes, Haptens, and Cross-Linking Reagents

Michael Brinkley

Molecular Probes, Inc., 4849 Pitchford Avenue, Eugene, OR 97402

Reprinted from *Bioconjugate Chemistry*, Vol. 3, No. 1, January/February, 1992

I. INTRODUCTION

Modification of proteins, DNA, and other biopolymers by labeling them with reporter molecules has become a very powerful research tool in immunology, histochemistry, and cell biology. A number of excellent reviews of this subject have been published (1-6). In addition, there are a growing number of commercial applications of these modified biomolecules, including clinical immunoassays, DNA hybridization tests, and gene fusion detection systems. In these techniques, a small molecule with special properties, such as fluorescence or binding specificity, is covalently bound to a protein, a DNA strand, or other biomolecule. Specific examples include fluorescently labeled antibodies for detection and localization of cell-surface antigens, biotin-labeled single-stranded DNA probes for detection of DNA hybridization, and hapten-labeled proteins that, when introduced into a suitable host animal, generate hapten-specific antibodies.

This review will focus on the experimental design and procedures for preparing protein conjugates with dyes, biotin, and haptens such as drugs and hormones. Methods for covalently linking two unlike biopolymers through the judicious choice of cross-linking reagents will also be discussed. The following specific topics will be addressed: (a) reactive groups of proteins that are available for modification, including their naturally occurring amino acids, and reactive groups introduced by chemical modification, (b) reagents that can be used to couple molecules to these reactive sites, (c) experimental procedures for preparing conjugates, (d) purification and isolation of conjugates, and (e) techniques for determining the degree of labeling.

II. GENERAL DISCUSSION OF METHODS

A. Reactive Groups of Proteins. Proteins and peptides are amino acid polymers containing a number of reactive side chains. In addition to, or as an alternative to, these intrinsic reactive groups, specific reactive moieties can be introduced into the polymer chain by chemical

modification. These groups, whether or not they are naturally a part of the protein or are artificially introduced, serve as "handles" for attaching a wide variety of molecules, including other proteins. The intrinsic reactive groups of proteins are described in the following section.

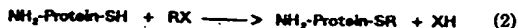
(1) *Amines (Lysines, α -Amino Groups)*. One of the most common reactive groups of proteins is the aliphatic ϵ -amine of the amino acid lysine. Lysines are usually present to some extent and are often quite abundant. For example, the protein bovine insulin contains only a single lysine amine, while avidin, a protein found in egg whites, contains 36 lysines (7). Lysine amines are reasonably good nucleophiles above pH 8.0 ($pK_a = 9.18$) (8) and therefore react easily and cleanly with a variety of reagents to form stable bonds (eq 1). Other reactive amines that are found



in proteins are the α -amino groups of the N-terminal amino acids. The α -amino groups are less basic than lysines and are reactive at around pH 7.0. Sometimes they can be selectively modified in the presence of lysines. There is usually at least one α -amino acid in a protein, and in the case of proteins that have multiple peptide chains or several subunits, there can be more (one for each peptide chain or subunit). Bovine insulin has one N-terminal glycine residue and one N-terminal phenylalanine (9). There are proteins that do not possess free α -amino groups, such as cytochrome C and ovalbumin. In these molecules, the N-terminal amino group is N-acylated, and therefore not reactive toward the usual modification reagents. Since either N-terminal amines or lysines are almost always present in any given protein or peptide, and since they are easily reacted, the most commonly used method of protein modification is through these aliphatic amine groups.

(2) *Thiols (Cystine, Cysteine, Methionine)*. Another common reactive group in proteins is the thiol residue from the sulfur-containing amino acid cystine and its reduction product cysteine (or half-cystine), which are counted together as one of the 20 amino acids. Cysteine contains a free thiol group, which is more nucleophilic

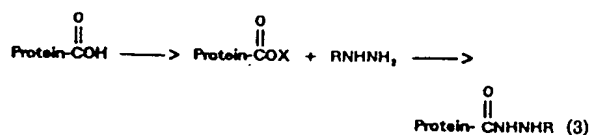
than amines and is generally the most reactive functional group in a protein. It reacts with some of the same modification reagents as do the amines discussed in the previous section and in addition can react with reagents that are not very reactive toward amines. Thiols, unlike most amines, are reactive at neutral pH, and therefore they can be coupled to other molecules selectively in the presence of amines (eq 2). This selectivity makes the thiol



group the linker of choice for coupling two proteins together, since methods which only couple amines (e.g., glutaraldehyde, dimethyl adipimide coupling) can result in formation of homodimers, oligomers, and other unwanted products (10). Since free sulfhydryl groups are relatively reactive, proteins with these groups often exist in their oxidized form as disulfide-linked oligomers or have internally bridged disulfide groups. Immunoglobulin M is an example of a disulfide-linked pentamer, while immunoglobulin G is an example of a protein with internal disulfide bridges bonding the subunits together. In proteins such as this, reduction of the disulfide bonds with a reagent such as dithiothreitol (DTT) is required to generate the reactive free thiol (11). In addition to cysteine and cysteine, some proteins also have the amino acid methionine, which contains sulfur in a thioether linkage. When cysteine is absent, methionine can sometimes react with thiol-reactive reagents such as iodoacetamides (12). However, selective modification of methionine is difficult to achieve and therefore is seldom used as a method of attaching small molecules to proteins.

(3) *Phenols (Tyrosine)*. The phenolic substituent of the amino acid tyrosine can react in two ways. The phenolic hydroxyl group can form esters and ether bonds, and the aromatic ring can undergo nitration or coupling reactions with reagents such as diazonium salts at the position adjacent to the hydroxyl group. There is considerable literature describing the reaction of tyrosyl residues with diazonium compounds (13). For example, a *p*-aminobenzoyl biocytin derivative has been diazotized and reacted with protein tyrosine groups (14). Modification of tyrosines has primarily been used in structural studies, rather than as a means for attaching specific labels, since acetylation and nitration can give useful information concerning the participation of tyrosine in the binding properties of proteins. Often, the reactivity of tyrosines with amine-selective modification reagents to form unstable carboxylic acid esters or sulfate esters is an unwanted side reaction resulting in conjugates that slowly hydrolyze during storage. Methods for preventing this problem are discussed in a later part of this teaching editorial (section V.B.1).

(4) *Carboxylic Acids (Aspartic Acid, Glutamic Acid)*. Proteins contain carboxylic acid groups at the carboxy-terminal position and within the side chains of the dicarboxylic amino acids aspartic acid and glutamic acid. The low reactivity of carboxylic acids in water usually makes it difficult to use these groups to selectively modify proteins and other biopolymers. In the cases where this is done, the carboxylic acid group is usually converted to a reactive ester by use of a water-soluble carbodiimide



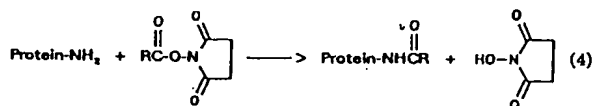
and then reacted with a nucleophilic reagent such as an amine or a hydrazide (15, 16). The amine reagent should be weakly basic in order to react specifically with the activated carboxylic acid in the presence of the other amines on the protein. This is because protein cross-linking can occur when the pH is raised to above 8.0, the range where the protein amines are partially unprotonated and reactive. For this reason, hydrazides, which are weakly basic, are useful in coupling reactions with a carboxylic acid (17). This reaction can also be used effectively to modify the carboxy terminal group of small peptides.

(5) *Other Amino Acid Side Chains (Arginine, Histidine, Tryptophan)*. Chemical modification of other amino acid side chains in proteins has not been extensive, compared to the groups discussed above. The high pK_a of the guanidine functional group of arginine ($pK_a = 12-13$) necessitates more drastic reaction conditions than most proteins can survive. Arginine modification has been accomplished primarily with glyoxals and α -diketone reagents (18). Tryptophan modification requires harsh conditions and is seldom carried out except as a method of analysis in structural or activity studies. Histidines have been subjected to photooxidation (19) and reaction with iodoacetates (20).

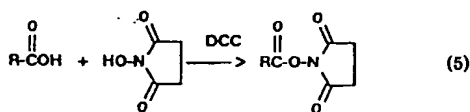
B. Protein Modification Reagents. This section will survey the extensive selection of reagents that are available for the purpose of protein modification. The fundamental principles for understanding how to use these reagents are (1) recognition of the reactive group(s) on the protein or peptide that can be modified and (2) knowledge of the type of chemical reactions these reactive groups will participate in and the nature of the chemical bonds that will result from these reactions.

(1) *Amine-Reactive Reagents*. These reagents are those which will react primarily with lysines and the α -amino groups of proteins and peptides under both aqueous and nonaqueous conditions. Some amine-reactive reagents are more reactive, and therefore less selective, than others, and it will be necessary to understand this property in order to choose the best reagent for modification of a specific protein. The following amine-reactive reagents are available.

(a) *Reactive Esters (Formation of an Amide Bond)*. Reactive esters, especially *N*-hydroxysuccinimide (NHS) esters, are among the most commonly used reagents for modification of amine groups (21). These reagents have intermediate reactivity toward amines, with high selectivity toward aliphatic amines. Their reaction rate with aromatic amines, alcohols, phenols (tyrosine), and histidine is relatively low. Reaction of NHS esters with amines under nonaqueous conditions is facile, so they are useful for derivatization of small peptides and other low molecular weight biomolecules. The optimum pH for reaction in aqueous systems is 8.0-9.0. The aliphatic amide products which are formed are very stable (eq 4). The

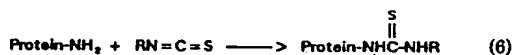


NHS esters are slowly hydrolyzed by water (22), but are stable to storage if kept well desiccated. Virtually any molecule that contains a carboxylic acid or that can be chemically modified to contain a carboxylic acid can be converted into its NHS ester (eq 5), making these reagents



among the most powerful protein-modification reagents available. Newly developed NHS esters are available with sulfonate groups that have improved water solubility (23). A short list of reactive NHS ester derivatives of fluorescent probes, biotin, and other molecules is given in Table I.

(b) *Isothiocyanates (Formation of a Thiourea Bond)*. Isothiocyanates, like NHS esters, are amine-modification reagents of intermediate reactivity and form thiourea bonds with proteins and peptides (eq 6). They are

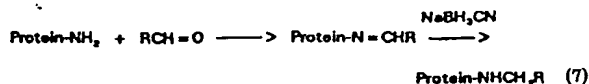


somewhat more stable in water than the NHS esters and react with protein amines in aqueous solution optimally at pH 9.0–9.5. Since this is a higher pH than the optimal pH for NHS esters (which undergo competing hydrolysis at pH 9.0–9.5), isothiocyanates may not be as suitable as NHS esters when modifying proteins that are sensitive to alkaline pH conditions. One of the most commonly used fluorescent derivatization reagents for proteins is fluorescein isothiocyanate (FITC). A number of other fluorescent dyes (coumarins and rhodamines) have been coupled to proteins via their reactive isothiocyanates (24).

(c) *Aldehydes (Formation of Imine, Reduction to Alkylamine Bond)*. Aldehyde groups react under mild aqueous conditions with aliphatic and aromatic amines to form an intermediate known as a Schiff base (an imine), which can be selectively reduced by the mild reducing agent sodium cyanoborohydride to give a stable alkylamine bond (eq 7) (44, 53). This method of amine modification is not used

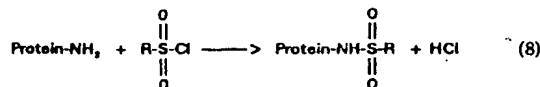
Table I. Succinimidyl Ester Probes

probes	structure	function	ref
succinimidyl fluorescein-5-(and -6)-carboxylate		fluorescent label	75, 76
succinimidyl <i>N,N,N',N'</i> -tetramethylrhodamine-5-(and -6)-carboxylate		fluorescent label	76
succinimidyl 7-amino-4-methylcoumarin-3-acetate		fluorescent label	77
succinimidyl X-rhodamine-5-(and -6)-carboxylate		fluorescent label	75, 78
succinimidyl D-biotin		ligand, affinity label	79
succinimidyl 3-(4-hydroxyphenyl)propionate		radioiodination label	80



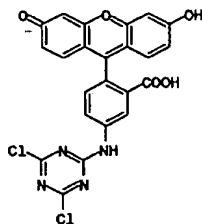
in protein conjugations as frequently as the activated ester method, but when the molecule to be attached has an aldehyde group, or can be easily converted to an aldehyde, the method is mild, simple, and very effective. Aldehydes (glyoxals) can also react with protein arginine groups (25, 26) and the nucleic acid base guanosine, making them of some use in nucleic acid modification (27).

(d) *Sulfonyl Halides (Formation of a Sulfonamide Bond)*. Sulfonyl halides are highly reactive amine-modifying reagents. They are unstable in water, especially at the pH required for reaction with aliphatic amines, but they form extremely stable sulfonamide bonds which can survive even amino acid hydrolysis (eq 8). It is for this



reason that sulfonamide conjugates are useful for amine-terminus derivatization (Dansyl-Edman degradation) and as tracers (28). In addition to amines, sulfonyl halides also react with phenols (tyrosine), thiols (cysteine), and imidazoles (histidine) on proteins (29); therefore, they are less selective than either NHS esters or isothiocyanates. The conjugates formed with thiols, imidazoles, and phenols are all unstable and, if not removed during purification, can lead to loss of the label from the protein during long-term storage (see section V.B.1). One of the most widely used long-wavelength fluorescent probes, Texas Red, is a sulfonyl chloride. It has the longest wavelength spectral properties of any of the common amine-reactive fluorescent labeling reagents (30).

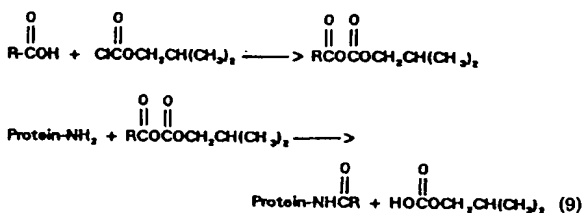
(e) *Miscellaneous Amine Reactive Reagents (Dichlorotriazines, Alkyl Halides, Anhydrides)*. The dichlorotriazine derivative of fluorescein, known as DTAF (I), has



I

high reactivity with protein amines and has been used to prepare fluorescein tubulin with minimal loss of activity (31). In addition to amines, dichlorotriazines will react with alcohols at elevated temperatures (60–90 °C) and are used to prepare polysaccharide conjugates (32). Some alkyl halides, including iodoacetamides commonly used to modify thiols, will react with amines of proteins if the pH is in the range 9.0–9.5 (33). Other reagents that have been used to modify amines of proteins are acid anhydrides. Succinic anhydride is commonly used to succinylate amine groups of basic proteins for the purpose of changing their isoelectric point and other charge-related properties (34). Mixed anhydrides derived from reaction of a carboxylic

acid with carbitol or 2-methylpropanol chloroformates (eq 9) are excellent reagents for modification of amines under

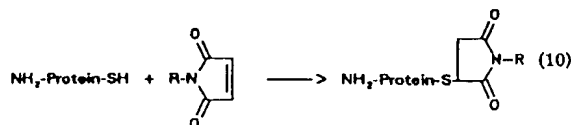


mild conditions (35). Of these, the carbitol mixed anhydride is relatively water soluble and is the preferred reagent for modification of amines in aqueous solution.

(2) *Thiol-Reactive Reagents*. Thiol-reactive reagents are those that will couple to thiol groups on proteins to give thioether-coupled products. These reagents react rapidly at neutral (physiological) pH and therefore can be reacted with thiols selectively in the presence of amine groups.

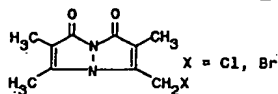
(a) *Haloacetyl Derivatives (Formation of a Thioether Bond)*. These reagents (usually iodoacetamides) are among the most frequently used reagents for thiol modification. In most proteins, the site of reaction is at cysteine groups that are either intrinsically present or that result from reduction of cystines. The reaction of iodoacetate with cysteine is approximately twice as fast as that with bromoacetate and 20–100 times as rapid as that with chloroacetate (36). As mentioned previously, in the absence of cysteines, methionines can sometimes react with haloacetamides (12). Reaction of haloacetamides with thiols occurs rapidly at neutral pH at room temperature or below, and under these conditions, most aliphatic amines are unreactive. In addition to proteins, haloacetamides have been reacted with thiolated peptides and thiolated primers for DNA sequencing (37), and also with RNA (on thiouridine) (38). The thioether linkages formed from reaction of haloacetamides are very stable. A potential problem in using iodoacetamides as modification reagents is their instability to light, especially in solution; therefore, they must be protected from light in storage and during reaction. The fluorescein and rhodamine iodoacetamides are among the most intensely fluorescent sulfhydryl reagents available for protein and peptide modification.

(b) *Maleimides (Formation of a Thioether Bond)*. Maleimides (eq 10) are similar to iodoacetamides in their



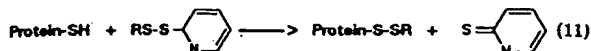
application as reagents for thiol modification; however, they are more selective than iodoacetamides, since they do not react with histidine, methionine, or thionucleotides (39, 40). The optimum pH for the reaction of maleimides is near 7.0. Above pH 8.0, hydrolysis of maleimides to nonreactive maleamic acids can occur (41).

(c) *Miscellaneous Thiol-Reactive Reagents*. These reagents include bromomethyl derivatives and pyridyl disulfides. The bromomethyl derivatives are similar in reactivity to iodoacetamides. The haloalkyl derivatives monobromobimane and monochlorobimane (II) react with



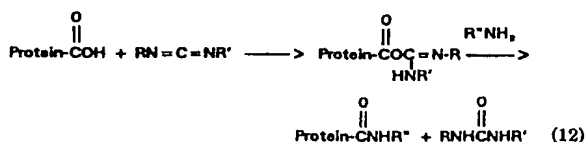
II

glutathione and other thiols in cells to give fluorescent adducts, thus providing a method of quantitation of thiols (42). Pyridyl disulfides react in an exchange reaction with protein thiols to give mixed disulfides (eq 11) (43).



(3) Carboxylic Acid- and Aldehyde-Reactive Reagents.

(a) Amines and Hydrazides (Formation of Amide or Alkylamine Bonds). Amines and hydrazides can be coupled to carboxylic acids of proteins via activation of the carboxyl group by a water-soluble carbodiimide followed by reaction with the amine or hydrazide. As mentioned previously (section II.A.4), the amine or hydrazide reagent must be weakly basic so that it will react selectively with the carbodiimide-activated protein in the presence of the more highly basic protein ϵ -amines (lysines). The reaction of these probes with carbodiimide-activated carboxyl groups leads to the formation of stable amide bonds (eq 12).



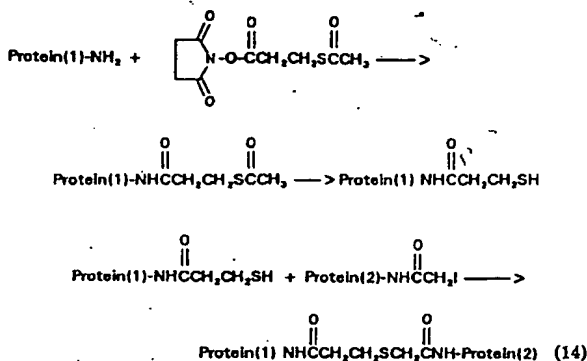
Amines and hydrazides are also able to react with aldehyde groups, which can be generated on proteins by periodate oxidation of carbohydrate residues on the protein. In this case, a Schiff base intermediate is formed (eq 13), which can be reduced to an alkylamine with sodium



cyanoborohydride, a mild and selective water-soluble reducing agent (44) (see also section II.B.1.c). Since the Schiff base formation is reversible, it is possible to minimize formation of protein-protein products by adding a large excess of amine or hydrazide reagent.

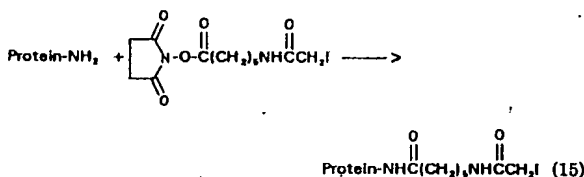
(4) Bifunctional Reagents. Bifunctional, or cross-linking, reagents are specialized reagents having reactive groups that will form a bond between two different groups, either on the same molecule or two different molecules. Bifunctional reagents can be divided into two types: those with the same reactive group at each end of the molecule (homobifunctional) and those with different reactive groups at each end of the molecule (heterobifunctional). Recent trends are heavily in favor of the use of heterobifunctional cross-linkers where the bifunctional reagent has two reactive sites, each with selectivity toward different functional groups (amine reactive and thiol reactive, for example). These reagents, some of which are available in a range of chain lengths, are well-suited to the task of controlled coupling of unlike biomolecules, such as two different proteins. Table II lists some frequently used heterobifunctional cross-linkers along with their reactivities and references describing their use.

(a) Amine Reactive—Thiol or Protected Thiol. Because thiols will react selectively in the presence of amines with a variety of reagents, these functional groups are very useful for attaching two different proteins together. Thiol-coupling methods are frequently employed to prepare protein-enzyme conjugates. If the proteins to be coupled do not contain intrinsic thiols, the procedure is typically carried out by introducing a single thiol group to an amine of one of the proteins by means of a heterobifunctional reagent (eq 14). Traut's reagent (iminothiolane) has been



extensively used for the purpose of introducing thiol groups selectively to proteins (45, 46). Many other bifunctional reagents contain both an amine-reactive and a protected thiol group, such as succinimidyl (acetylthio)acetate (SATA) (47, 48) or succinimidyl 3-(2-pyridyldithio)propionate (SPDP) (43, 49). After deprotection, the thiol-containing protein is then reacted with a thiol-reactive group on the other protein, which has been introduced by a similar technique. Alternatively, proteins with synthetic thiol groups that have been introduced by modification can be used to couple to a number of thiol-reactive derivatives of dyes, biotin, haptens, or other molecules.

(b) Amine Reactive—Iodoacetamide. Iodoacetamides are primarily thiol-reactive groups with the reaction occurring rapidly at physiological pH, but they can react with amines under more alkaline conditions (greater than pH 9.0) and long reaction times (section II.B.2.a). Iodoacetamides can be introduced into a protein or peptide that does not have intrinsic thiols via amine-reactive derivatives (eq 15) (50). The resulting modified protein



can then be coupled to any thiol-containing molecule. The second molecule is usually a thiol-containing protein.

(c) Amine Reactive—Maleimide. The introduction of maleimides into a protein or peptide can be carried out with heterobifunctional reagents that have an amine-reactive group at one end and the thiol-specific maleimide at the other end (eq 16). The applications are very

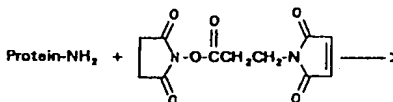
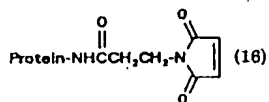


Table II. Heterobifunctional Cross-Linking Reagents

reagent	structure	reactivity	ref
succinimidyl 3-(2-pyridyldithio)propionate (SPDP)		primary amine, thiol	49
succinimidyl <i>trans</i> -4-(<i>N</i> -maleimidylmethyl)cyclohexane-1-carboxylate (SMCC)		primary amine, thiol	54, 48
succinimidyl (acetylthio)acetate (SATA)		primary amine, thiol	47, 48
4-[(succinimidylthio)carboxyl]- α -methyl-2-pyridyldithio]toluene (SMPT)		primary amine, thiol	55, 48
succinimidyl 4-[[[iodoacetyl]amino]methyl]-cyclohexane-1-carboxylate (SIAC)		primary amine, thiol	50
succinimidyl <i>p</i> -azidobenzoate (SAB)		primary amine, nonselective	56



similar to those for the iodoacetamides discussed in the preceding section. Specific applications include coupling of ricin to monoclonal antibodies (51) and linking of oligonucleotides to enzymes (52).

(d) *Amine Reactive—Aldehyde*. Aldehydes do not occur naturally in proteins, but can be introduced in two ways. In the first method, carbohydrate groups on proteins are treated with an oxidizing reagent, such as sodium periodate, or are converted via a galactose oxidase/catalase enzyme method, both of which split the sugar to form aldehyde groups (53). Not all proteins contain carbohydrate groups, and therefore a second method of introducing aldehydes via the reagent glutaraldehyde has been employed (10). Glutaraldehyde has been used extensively to couple two proteins together via their amine groups (eq 17); however, like other homobifunctional reagents, glu-



taraldehyde is being replaced with more selective heterobifunctional reagents such as those discussed above.

(5) *Photoactivatable Reagents*. Reagents are available that can be activated by light (photons) to produce a reactive intermediate that can couple to various functional

groups on biomolecules. Two of the most frequently used photoactivatable reagents for this purpose are aromatic azides and benzophenones.

(a) *Aromatic Azides*. Aromatic azides are efficiently photolyzed by illumination with an ultraviolet light at 300–350 nm. The reactive molecule produced by this photolysis is a nitrene, which reacts rapidly and nonspecifically with either solvent molecules or with functional groups on biomolecules. Almost any functional group or amino acid can be modified, since the nitrene is very reactive. Recent improvements in azide-based protein modification reagents have resulted in perfluorinated azides that generate nitrene intermediates with greater stability, thus giving reagents with higher efficiency (up to 40%) of reaction with the protein (57, 58). One of the primary uses of these highly reactive reagents is to carry out photoaffinity labeling experiments. In these experiments, the aromatic azide is attached to a drug or other molecule which binds specifically to a protein binding site (an example is an enzyme inhibitor or a nucleotide analogue) and then photolyzed. The location and type of bond formed in this process provides information about the environment near the binding site (59). In addition to their role as photoaffinity labels, aryl azides are useful as heterobifunctional cross-linkers. Succinimidyl azidobenzoate (SAB), *p*-azidophenacyl bromide, and 4-maleimidobenzophenone have been employed to couple proteins through dark reaction with amines or thiols followed by light activation (56, 58, 60, 61).

(b) *Benzophenones*. Benzophenones are like azides in that they are photoactivatable by ultraviolet light, but

once they have been activated, they can either react with functional groups or return to the ground state. Thus, these molecules can sometimes be reactivated if they do not react on the first activation. These reagents are also used as photoaffinity labels in a manner similar to that of the aromatic azides (62).

III. PRACTICAL CONSIDERATIONS

Along with a thorough knowledge of protein reactivity and the available reagents for the desired type of protein modification, it is of crucial importance that the researcher understand the practical aspects of carrying out reactions between highly reactive small organic molecules and large, complex, conformationally sensitive, water-soluble biopolymers. The following discussion will address some of the general rules, problems, and pitfalls of protein-modification chemistry.

A. Choosing the Right Buffer. Conjugations should be carried out in a well-buffered system at a pH that is optimal for the reaction. The ionic strength should, in most cases, be in the range of 25–100 mM. For modification of thiol groups and α -amino groups, which occurs selectively at physiological pH (7.0–7.5), phosphate buffers are ideally suited. The more strongly basic lysine amines require more alkaline pH, in the range of 8.0–9.5, where phosphate solutions do not buffer well. For these reactions, carbonate/bicarbonate (pH of 100 mM bicarbonate is 9.2) or borate buffers are quite satisfactory. As an example, conjugations with NHS esters are best carried out in pH 8.2 bicarbonate buffer, while isothiocyanates require the higher pH (9.0–9.5) provided by carbonate or borate buffers. The choice of buffer will in some cases be directed by compatibility of the protein.

B. Cosolvents. If the reagent that is to be attached to the biomolecule is readily soluble at millimolar concentrations in water or buffer, no cosolvent is needed, and the reagent can be added as a concentrated aqueous solution to the buffered reaction solution. Unfortunately, aqueous systems are very often incompatible with the reagent, as a result of poor solubility or high reactivity with water. In these cases, a water-miscible cosolvent must be employed that will dissolve the reagent without causing its decomposition. At the same time, the cosolvent must not cause irreversible denaturation or precipitation of the biomolecule. Some cosolvents that have been successfully utilized in protein modifications are methanol, ethanol, 2-propanol, 2-methoxyethanol, dioxane, dimethylformamide (DMF), and dimethyl sulfoxide (DMSO).

The most versatile of these cosolvents are DMF and DMSO. They are recommended because of the following desirable properties: (a) they are inert to many of the reactive reagents used in preparing conjugates, (b) they are miscible with water in all proportions, and (c) they are compatible with most aqueous protein solutions even at up to 30% v/v ratios. DMF is the solvent of choice for reactions of sulfonyl chlorides, since these reagents will react with DMSO. It is usually important that cosolvents be carefully dried and stored over a drying agent to prevent competing hydrolysis of the reactive modification reagent.

C. Reaction Conditions. As a general rule, conjugation reactions should be done at below room temperature, since the rate of reaction of most conjugation reagents is rapid at low temperature. Low temperatures tend to increase the selectivity of the reaction, resulting in fewer side reactions and more consistent and reproducible results. A convenient procedure is to add the reagent to a gently stirred buffered solution of the protein in an ice-bath and then allow the bath to warm to room temperature over a

period of about 2 h. Very reactive reagents such as sulfonyl chlorides should be reacted under more carefully controlled conditions, such as 4 °C for 1 h. Stirring can be done with a magnetic stir-bar and should not be excessively fast, since proteins can be denatured by violent mixing. Addition of the reagent should be carried out dropwise and as slowly as possible, since gradual addition increases the selectivity of the reaction.

(1) *Protein Concentration.* Because the kinetics of conjugation of these reagents is bimolecular, but the hydrolysis rates are pseudo-first-order, dilution results in competition between conjugation and loss of reagent by hydrolysis. Protein concentrations above 10 μ M are strongly recommended, with an optimum in the range of 50–100 μ M.

(2) *pH.* In modification of amines, only the unprotonated form is reactive, and therefore it is necessary to maintain a pH at which a significant number of amines are unprotonated. An average pK_a above 9 for lysines indicates that the higher the pH, the better. Offsetting this are the factors that the rate of reagent hydrolysis increases rapidly above pH 9 and that proteins tend to be unstable at a higher pH. A free amine terminus has a pK_a near 7 and is sometimes preferentially modified when the reaction is run at neutral pH. An effective compromise in most cases is to use a pH close to 9.0–9.2 if the protein is stable, but a lower pH combined with more reagent and longer reaction times if the protein is unstable. The succinimidyl esters and DTAF appear to react more efficiently at a lower pH than the isothiocyanates and sulfonyl chlorides. Our experience with succinimidyl esters indicates that a reaction pH of around 8.2 gives excellent results for most proteins.

(3) *Reaction Time.* Usually, 1–2 h is sufficient time for conjugation reactions to go to completion. Longer reaction times, if convenient, are acceptable, since the degree of labeling is generally limited by the ratio of the reagent to protein, rather than the reaction time. Many published procedures specify overnight reaction times. Obviously, the more reactive the reagent, the shorter the reaction time; sulfonyl chloride reactions are faster than NHS ester reactions.

IV. FACTORS INFLUENCING CHOICE OF MOLAR RATIO OF REACTANTS

A. End Use of Reagents. (1) *Immunogen—High Degree of Labeling.* Protein conjugates are frequently prepared for use in producing specific antibodies to a drug or other hapten in a host animal. The drug or hapten is conjugated to a high molecular weight protein carrier molecule and injected into the animal to elicit an immune response, and over a period of time, specific antibodies to the drug or hapten are produced. For these purposes, a high degree of labeling of the protein carrier is desirable, since more labels generally increase the strength and specificity of the immune response.

(2) *Labeled Antibody or Enzyme—Low to Moderate Degree of Labeling.* Antibodies and enzymes are relatively sensitive to substitution, since there are usually reactive amino acid side chains (amines, thiols, histidines) in or near the binding sites. For this reason, a low to moderate degree of labeling is preferred in order to preserve binding specificity or enzyme activity. Excessive labeling can also result in decreased solubility of the conjugates, which also reduces the overall activity. In the case of many fluorescent labels, a high dye to protein ratio causes a dramatic decrease in the fluorescence efficiency of the conjugates

(63, 64). In our experience with antibodies, a substitution ratio in the range of 4–6 is usually optimal for good retention of binding activity.

(3) *Fluorescent Labeled Proteins/Peptides—Low to Moderate Degree of Labeling.* Fluorescent labels are often very sensitive to their molecular environment and therefore their fluorescence intensity is almost always decreased when they are bound to proteins and other biomolecules. Fluorescence also decreases when the fluorescent labels are located in close proximity to one another, probably as a result of transfer of excited-state energy (quenching) from one molecule to another (65). When proteins are labeled with fluorescent dyes, the fluorescence increases as more dyes are added; at the same time, however, the fluorescence efficiency decreases as a result of the quenching described above. Some dyes are more sensitive to quenching than others. FITC is about 50–70% quenched on IgG at a dye/protein ratio of 5 (66), while Cascade Blue, a newly developed blue fluorescent dye (67), retains nearly 100% of its fluorescence efficiency under the same conditions. The number of dyes that can be conjugated to a protein without substantial loss of fluorescence will depend on the size of the protein and the distance between the functional groups to which the label is attached. Usually, more dyes can be attached to a large protein than a small protein or peptide. A general rule for conjugates of fluorescein is 4–6 dyes/protein and for rhodamines, 2–3 dyes/protein. The degree of labeling depends on the relative reactivity of the labeling reagent to the protein and to water, the molecular weight and number of reactive amines on the protein, the reactant concentrations (especially of the protein), and other factors. The exact amount of label to use must be determined by experiment; however, as a guideline, 10 mol of a typical isothiocyanate or NHS ester is needed to label 1 mol of a protein. Because of the faster competitive hydrolysis rate, 20 mol of a sulfonyl chloride, such as Texas Red, is required to label 1 mol of a protein.

B. Number of Reactive Groups on the Protein. Proteins vary greatly in the number of reactive amino acid groups. For example, some proteins have 40 or more reactive amine groups, while others may have only one or two amines or thiol groups. The reactivity of these groups with the labeling reagent and their effective concentration in solution will then have an effect on the amount of labeling reagent required to achieve the desired degree of substitution. This means that small molecular weight proteins or peptides with few reactive groups will require more labeling reagent per gram than large molecular weight proteins with many reactive groups.

C. Solubility of Modification Reagent in Reaction Solution. (1) *Cosolvent Sometimes Required.* The use of cosolvents was explained in section III.B. In some cases the labeling reagent is very hydrophobic and, even though it is readily soluble in DMF or DMSO, it precipitates when added to the buffered protein solution. It is often possible to circumvent this problem by adding some cosolvent gradually, with stirring, to the buffered protein solution until the protein solution contains 20–25% cosolvent. The ionic strength of the buffer should be no more than 50 mM so that the buffer does not salt out upon addition of the cosolvent. Then the solution of labeling reagent in cosolvent is added so that the final volume percent cosolvent in the reaction mixture is around 30%. This modification often is successful in preventing precipitation of the labeling reagent. Many proteins are stable in 30% DMSO or DMF; however the stability of the protein to

these conditions should be determined before carrying out this technique.

(2) *Two-Stage Labeling as a Last Resort.* If the technique described in section IV.C.1 is used and the labeling reagent still precipitates when added to the protein solution, it may be possible to purify the conjugate and then repeat the labeling procedure to increase the degree of substitution.

D. Solubility of Conjugate. (1) *Conjugate Is Often Less Soluble Than Native Protein.* Problems with solubility of the conjugate can occur, most often when the labeling reagent is hydrophobic or contains multiple ionic groups. These physical properties of the label can upset the natural folding of the protein and cause the conjugate to be significantly less soluble than the native protein (30).

(2) *Overlabeling Can Cause Precipitation of Conjugate.* Overlabeling can produce the same undesirable results noted above. The best solution to these problems is to use a lower ratio of labeling reagent to protein, resulting in a conjugate with a lower degree of substitution.

V. PURIFICATION OF CONJUGATES

A. Removal of Excess Noncovalently Bound Labeling Reagent. (1) *Dialysis—Simple, Inexpensive Purification Method—Inefficient for Hydrophobic Molecules.* Dialysis is the simplest, but most time-consuming, method of purifying protein conjugates. Not all molecules dialyze efficiently; the rate of dialysis depends on their relative affinity for the protein versus the dialysis solution. Molecules that are sparingly soluble in water or strongly adsorbed to the protein surface will take a long time to dialyze. Dialysis works best when the labeling reagent and its unreacted byproducts are hydrophilic. When purifying conjugates by dialysis, a dialysis buffer volume of at least 100 times the volume of the conjugate solution should be used and the dialysis buffer should be changed at least five times. Allow at least 4 h for dialysis between buffer changes.

(2) *Gel Filtration—Faster Than Dialysis—Effectively Removes Most Hydrophilic and Hydrophobic Labeling Reagents.* Gel exclusion chromatography separates conjugates from excess noncovalently bound labeling reagent and other small molecular weight impurities by selectively adsorbing the small molecules, while allowing the larger protein conjugate molecules to pass through the void space in the gel. This method is very fast and effective for purifying conjugates from both hydrophobic and hydrophilic labeling reagents. A common technique employs a Sephadex G-25 or similar column containing about a 2-mL bed volume/mg of protein that can be packed in any suitable buffer (30). Upon elution in the case of dyes, the conjugate and free dye bands are usually clearly visible; many other types of labels can be visualized by holding a hand-held UV lamp close to the column during chromatography. Automatic fraction-collecting devices with UV monitors are also frequently used. If partial precipitation has occurred during the reaction, the samples should be centrifuged before running the column. The solution of labeled protein will contain a mixture of species with variable degrees of substitution. If required, separation of the lightly and heavily labeled fractions can be done by ion-exchange chromatography. Usually one passage through a gel filtration column is sufficient to remove most of the unreacted label; however, some proteins bind small molecules with high avidity. To completely

purify these conjugates it may be necessary to carry out additional purification steps.

(3) **Hydrophobic Interaction Adsorbents—Removes Strongly Bound Hydrophobic Labeling Reagents.** Some labeling reagents have a very strong affinity for certain proteins and cannot be completely removed by gel filtration. These conjugates can be further purified (after gel filtration to remove most of the unreacted label) by treatment with microporous, hydrophobic polystyrene beads (68). In this procedure, the conjugate is simply mixed with the beads, and the small hydrophobic molecules are selectively adsorbed into the micropores while the larger conjugate molecules are excluded.

B. Removal of Labeling Reagent Attached by Unstable Covalent Bonds. (1) **Hydroxylamine Treatment—Hydrolysis of Tyrosine Ester Bonds under Mild Conditions.** Section II.A.3 describes the formation of tyrosine esters. Several of the reagents commonly used for protein modification, including NHS esters, isothiocyanates, and sulfonyl chlorides, can react with tyrosines to form these esters. These adducts are unstable and can hydrolyze even at physiological pH, resulting in loss of label over a period of time. Since any measurable loss of label can interfere with the intended use of many conjugates, it is advisable to pretreat all conjugates prepared with these types of reagents to remove any esters that may have formed in the conjugation reaction. This can be effectively done in most cases by treating the conjugate before purification with hydroxylamine (69, 70). In this method, a 1.5 M solution of hydroxylamine at pH 8.0 is added to the conjugate solution to a final concentration of 0.1 M and the solution is stirred at room temperature for 1 h. The conjugate is then purified by gel filtration or dialysis.

VI. EXPERIMENTAL METHODS FOR PREPARING PROTEIN CONJUGATES

The general experimental procedures that follow describe methods for conjugating amine-reactive and thiol-reactive probes to proteins. They should be useful as a guide for the experimentalist; however, it is strongly suggested that the numerous literature references given in this review and others be consulted for additional specific information. Because of the very wide variety of experimental conditions required for coupling proteins with bifunctional reagents, it is difficult to generate a simple general procedure and the reader is advised to consult the literature for specific procedures.

A. Amine-Reactive Probes. The following general procedure is recommended for the first trial and is adaptable to amine-reactive dye, biotin, hapten, and bifunctional linker conjugations. The procedure may be modified after the degree of substitution has been determined (see below) after purification.

Step 1. Dissolve the protein at 50–100 μ M in 50–100 mM sodium bicarbonate buffer at pH 9.2 at room temperature. Borate buffer is also suitable. Amine-based buffers, such as TRIS are not recommended. Conjugations with succinimide esters and reagents such as DTAF [5-[(4,6-dichlorotriazin-2-yl)amino]fluorescein] should be done at a lower pH. In these cases, a suitable buffer is 50–100 mM pH 8.2 sodium bicarbonate.

Step 2. Add sufficient protein-modification reagent from a stock solution to contain about 10 mol of isothiocyanate or succinimide ester for each mole of protein or about 20 mol of sulfonyl chloride for each mole of protein.

Although most protein modification reagents have some solubility in water, it is recommended that a stock solution be prepared immediately before use in a water-miscible nonhydroxylic solvent such as dimethyl formamide (DMF), dimethyl sulfoxide (DMSO), or dioxane. The stock solution should be prepared fresh each time, since it is very difficult to store these solutions for any length of time without decomposition of the reagent taking place. As a guideline, it is recommended to prepare a stock solution at about 10–20 mM of the protein-modification reagent in dry DMF. The fluorescent dyes Texas Red, Lissamine rhodamine B, and other sulfonyl chlorides must never be used in DMSO, with which they react. These stock solutions (prepared in dry DMF) are usually diluted about 10-fold into the protein, while being agitated to avoid high local concentrations of reagent. Some reagents are quite hydrophobic, having little solubility in the aqueous protein solution. This is particularly true of some of the rhodamine and biotin succinimidyl esters. A technique that helps in these cases is to add a 20% volume of DMF or DMSO slowly to the protein/buffer solution before adding the stock solution of the reagent in DMF or DMSO (see section IV.C.1).

Isothiocyanates and Succinimidyl Esters. Add the solution of the modification reagent dropwise using a microliter syringe during a period of about 1 min to the stirred protein solution while in an ice-water bath. Allow the reaction mixture to warm to room temperature and continue to stir for at least 2 h.

Sulfonyl Chlorides. Add the solution of the reagent quickly using a micropipet to the stirred protein solution in an ice bath or in a cold room. Allow to react at 4 °C for 1 h.

Step 3. Separate the conjugate from unreacted dye on a gel filtration column using the appropriate buffer as described in section V. Texas Red and certain other rhodamine-based conjugates will still retain varying amounts of noncovalently adsorbed dye even after purification by gel chromatography. This protein-adsorbed dye can be removed by treating the conjugate with a hydrophobic adsorbent as described in section V.A.3.

B. Thiol-Reactive Probes. A general procedure suitable for conjugation of thiol-reactive probes, including maleimides, iodoacetates, and alkyl halides, is outlined below. As a rule, thiol-reactive reagents are more stable to water than the reactive esters; however, they should be handled carefully and stored in a freezer with protection from light and moisture. As with the reactive esters and isothiocyanates discussed above, only freshly prepared reagent solutions should be used. Protection from light is particularly important for iodoacetamides.

Step 1. Dissolve the protein at 50–100 μ M in a suitable buffer at pH 7.0–7.5 (10–100 mM phosphate, TRIS, HEPES) at room temperature. At this pH range, the protein thiol groups are sufficiently nucleophilic so that they react almost exclusively with the reagent in the presence of the more numerous protein amines, which are protonated and relatively unreactive. As a general rule, it is advisable to carry out thiol modifications in an oxygen-free environment, since some thiols can be oxidized to disulfides. This is particularly important if the modification reagent is to be reacted with a cystine group that has been previously reduced with a reagent such as dithiothreitol. In this case, all buffers should be deoxygenated and the reactions carried out under an inert atmosphere to prevent re-formation of disulfide.

Step 2. Add sufficient protein modification reagent from a stock solution of the reagent to contain 10–20 mol of reagent for each mole of protein. If the reagent is water-soluble, an aqueous solution can be used; otherwise, the reagent can be dissolved in one of the water-miscible non-hydroxylic solvents recommended for use with amine-reactive reagents. The reagent concentration should be about 10–20 mM. Upon completion of the reaction with the protein, an excess of glutathione, mercaptoethanol, or other soluble low molecular weight thiol can be added to consume excess modification reagent, thus ensuring that no reactive species are present during the purification step.

Iodoacetamides. Reactions with iodoacetamides should be carried out in the dark, since light can cause reagent decomposition. Add the stock reagent solution dropwise and slowly to the gently stirred solution of the protein at room temperature over a period of about 1 min. Continue stirring for 2 h.

Maleimides. Reaction conditions are essentially the same as with iodoacetamides; however, the selectivity of maleimides toward thiol groups is greater, allowing somewhat more latitude in the buffer pH. Decomposition to maleamic acids above pH 8.0 is a competing reaction. Add the stock reagent solution dropwise and slowly to the gently stirred protein solution at room temperature over a period of about 1 min and allow the mixture to react for 2 h.

Step 3. Separate the conjugate from unreacted modification reagent as described in section V.

C. Storage of Conjugates. Conjugates should be stored as one normally stores the parent protein. If the protein is stable to freezing, then lyophilization is recommended for long term storage. Sodium azide at 2 mM or thimerosal may be added to inhibit bacterial growth. **CAUTION:** These preservatives may be toxic in live-cell use of conjugates. In addition, sodium azide is an inhibitor of the enzyme horseradish peroxidase (HRP). Therefore, thimerosal should be substituted as a preservative in situations where the conjugate is derived from HRP or it is anticipated that the conjugate will be used in the presence of HRP. Fluorescent dye conjugates should be protected from light.

VII. DETERMINATION OF THE DEGREE OF SUBSTITUTION OF PROTEIN CONJUGATES

Several methods are available for determining the degree of substitution of modified proteins. If the modification results in the creation of thiol residues, as is often the case with bifunctional reagents, it is relatively straightforward to determine the degree of substitution by quantitation of thiols. Several colorimetric methods for thiol determination are available (43, 45, 47). Maleimides introduced into proteins can be determined by back-titration with 2-mercaptoethanol (81). Dyes and many other types of molecules introduced into proteins are usually determined by spectroscopic techniques, as described below.

This general procedure should be applicable to dyes and other molecules that have significant absorption above 280 nm.

The determination of dye/protein (D/P) levels by spectroscopy is accomplished by determining the apparent concentration of dye in the conjugate by measuring its absorption at its characteristic λ_{max} and then measuring the protein concentration of the conjugate by its absorption at 280 nm. Because most dyes have some absorption at 280 nm, the absorption of the conjugate at 280 nm must be corrected for the contribution of the dye to obtain the correct protein concentration. The ratio of these two

concentrations, calculated by use of Beer's law ($A = \epsilon Cl$, where ϵ = extinction coefficient, A = molar absorbance, C = molar concentration, and l = path-length), is then equal to the D/P ratio.

This method is inexact, because there is no way to know precisely how the spectral characteristics of the dye change when it is conjugated to the protein. The following assumptions and approximations are made.

(1) The extinction coefficient of the protein-bound dye at its absorption maximum is about the same as the extinction coefficient of the free dye in solution at its absorption maximum. Although there are undoubtedly some differences, experiments have shown that this assumption is at least approximately correct (64).

(2) The absorption of the protein-bound dye at 280 nm is about the same as the absorption of the free dye in solution. This assumption may be less reliable than the previous assumption, since there is probably more contribution from the linking group to this portion of the spectrum, and this group can be substantially changed when attached to the protein. The following question arises: what is the "free dye"? There is no unambiguous answer to this question, since the dye, when attached to the protein, is different than the free dye, and the spectral properties will be somewhat different. The best choice of free dye if the NHS ester was used as the reagent is probably the free acid or lysine amide derivative. These may be available or can be synthesized. *Do not use the NHS ester as the free dye, since the N-succinimidyl group absorbs strongly at 280 nm.* In other cases, sulfonic acids can be used when the protein modification reagent was a sulfonyl chloride.

(3) The extinction coefficient of the conjugate at 280 nm is about the same as the extinction coefficient of the native protein. However, extensive modification of the protein may change the spectral absorption at 280 nm in an unknown manner.

Although there are obvious questionable assumptions, spectroscopy remains the easiest and most convenient method of determining D/P ratios. One alternative is to determine the protein concentration by weighing the conjugate, which eliminates problems in assumption 3, but this is tedious and includes the danger that the conjugate will denature when dried without buffer, or the lyophilized conjugate may contain entrapped buffer salts. This method does not eliminate errors from assumptions 1 and 2. Another alternative is to digest a known amount of the conjugate chemically or with a proteolytic enzyme to degrade the molecule to small fragments containing the dye and then determine the concentration of the dye by spectroscopy. This is even more tedious and still does not usually give a pure dye product which can be compared spectrally with a known derivative. Because of the lack of convenient and suitable alternatives, direct spectroscopic determination is the most frequently used method of estimating D/P ratios (64, 71–74).

Procedure. *Step 1.* Obtain absorption spectra of the free dye and the dye-protein conjugate (note 1).

Step 2. Obtain extinction coefficients of the free dye and protein from a handbook of dyes and protein tables (8, 50).

Step 3. Perform these calculations:

$$C_d = A_{\lambda_{\text{max}}} / \epsilon_d$$

$$F = A_{d(280)} / A_d$$

$$C_p = [A_{280} - (A_{\lambda_{\max}} F)] / \epsilon_p$$

$$D/P = C_d / C_p$$

where ϵ_d is the extinction coefficient of free dye at λ_{\max} , A_d is the absorbance of free dye at λ_{\max} , $A_{d(280)}$ is the absorbance of free dye at 280 nm, $A_{\lambda_{\max}}$ is the absorbance of dye in conjugate at λ_{\max} , ϵ_p is the extinction coefficient of protein at 280 nm, A_{280} is the absorbance of protein in conjugate at 280 nm, C_d is the concentration of dye in conjugate (mol/L), and C_p is the concentration of protein in conjugate (mol/L).

ACKNOWLEDGMENT

I thank Dr. Rosaria Haugland and Danuta Szalecki for helpful discussions and advice concerning experimental details. I also thank Nan Minchow for preparing the structures and tables.

LITERATURE CITED

- (1) (a) Means, G. E., and Feeney, R. E. (1971) *Chemical Modification of Proteins*. Holden-Day, San Francisco, CA. (b) Means, G. E., and Feeney, R. E. (1990) *Chemical Modification of Proteins: History and Applications*. *Bioconjugate Chem.* 1, 2.
- (2) Glazer, A. N., Delange, R. J., and Sigman, D. S. (1975) *Chemical Modification of Proteins. Laboratory Techniques in Biochemistry and Molecular Biology* (T. S. Work, and E. Work, Eds.) American Elsevier Publishing Co., New York.
- (3) Lundblad, R. L., and Noyes, C. M. (1984) *Chemical Reagents for Protein Modification*, Vols. I and II, CRC Press, New York.
- (4) Pfeleiderer, G. (1985) *Chemical Modification of Proteins. In Modern Methods in Protein Chemistry* (H. Tschesche, Ed.) Walter DeGruyter, Berlin and New York.
- (5) Eyzaguirro, J. (1987) *Chemical Modification of Enzymes. Active Site Studies*. John Wiley & Sons, New York.
- (6) Wong, S. H. (1991) *Chemistry of Protein Conjugation and Cross-linking*, CRC Press, Boca Raton, FL.
- (7) De Lange, R. J., and Huang, T. S. (1971) Egg white avidin. III. sequence of the 78-residue middle cyanogen bromide peptide. Complete amino acid sequences of the protein subunit. *J. Biol. Chem.* 246, 698.
- (8) Fasman, G. D., Ed. (1989) *Practical Handbook of Biochemistry and Molecular Biology*, p 13, CRC Press, Boca Raton, FL.
- (9) White, A., Handler, P., and Smith, E. L. (1982) *Principles of Biochemistry*, p 142, McGraw-Hill, New York.
- (10) (a) Korn, A. H., Fairheller, S. H., and Filachione, E. M. (1972) Glutaraldehyde: nature of the reagent. *J. Mol. Biol.* 66, 525. (b) Hardy, P. M., Nicholls, A. C., and Rydon, N. H. (1976) The nature of the crosslinking of proteins by glutaraldehyde. Part 1. Interaction of glutaraldehyde with the amino group of β -aminohexanoic acid and of α -N-acetyl-lysine. *J. Chem. Soc. Perkin Trans. 1*, 958.
- (11) Cleland, W. W. (1964) Dithiothreitol, a new protective reagent for SH groups. *Biochemistry* 3, 480.
- (12) Gundlach, H. G., Moore, S., and Stein, W. H. (1959) The reaction of iodoacetate with methionine. *J. Biol. Chem.* 234, 1761.
- (13) Riordan, J. F., and Vallee, B. L. (1972) Diazonium salts as specific reagents and probes of protein configuration. *Methods Enzymol.* 25, 251.
- (14) Wilchak, M., Ben-Hur, H., and Bayer, E. A. (1966) p-Di-azobenzoyl biocytin—A new biotinylating reagent for the labelling of tyrosines and histidines in proteins. *Biochem. Biophys. Res. Commun.* 136, 872.
- (15) Hoare, D. G., and Koshland, D. E., Jr. (1966) A procedure for the selective modification of carboxyl groups in proteins. *J. Am. Chem. Soc.* 88, 2087.
- (16) Yamada, H., Imoto, T., Fujita, K., Ozaki, K., and Motomura, M. (1981) Selective modification of aspartic acid 101 in lysozyme by carbodiimide reaction. *Biochemistry* 20, 4836.
- (17) Renthall, R., Cothran, M., Dawson, N., and Harris, G. J. (1987) Fluorescent labeling of bacteriorhodopsin: implications for helix connections. *Biochim. Biophys. Acta* 897, 384.
- (18) Yankeelov, J. A., Jr., Mitchell, C. D., and Crawford, T. H. (1968) A simple trimerization of 2,3-butanediones yielding a selective reagent for the modification of arginine in proteins. *J. Am. Chem. Soc.* 90, 1664.
- (19) Bond, J. S., Francis, S. H., and Park, J. H. (1970) An essential histidine in the catalytic activities of 3-phosphoglyceraldehyde dehydrogenase. *J. Biol. Chem.* 245, 1041.
- (20) Stark, G. R., Stein, W. H., and Moore, S. (1961) Relationships between the conformation of ribonuclease and its reactivity toward iodoacetate. *J. Biol. Chem.* 236, 436.
- (21) Bragg, P. D., and Hou, C. (1975) Subunit composition, function and spatial arrangement in the Ca^{2+} and Mg^{2+} -activated adenosine triphosphatases of *Escherichia coli* and *Salmonella typhimurium*. *Arch. Biochem. Biophys.* 167, 311.
- (22) Lomants, A. J., and Fairbanks, G. (1976) Chemical probes of extended biological structures: synthesis and properties of the cleavable protein cross-linking reagent [^{35}S]dithiobis(succinimidyl propionate). *J. Mol. Biol.* 104, 243.
- (23) Staros, J. V. (1982) *N*-hydroxysulfosuccinimide active esters: Bis(*N*-hydroxysulfosuccinimide) esters of two dicarboxylic acids are hydrophilic, membrane-impermeant protein cross-linkers. *Biochemistry* 21, 3950.
- (24) Brantzag, P. (1975) Rhodamine conjugates: specific and non-specific binding properties in immunohistochemistry. *Ann. N.Y. Acad. Sci.* 254, 35.
- (25) Takihashi, K. (1968) The reaction of phenylglyoxal with arginine residues. *J. Biol. Chem.* 243, 6171.
- (26) Konishi, K., and Fujioka, M. (1987) Chemical modification of a functional arginine residue of rat liver glycine methyltransferase. *Biochemistry* 26, 8496.
- (27) Wagner, R., and Gassen, H. G. (1975) On the covalent binding of mRNA models to the part of the 16S RNA which is located in the mRNA binding site of the 30S ribosome. *Biochem. Biophys. Res. Commun.* 65, 519.
- (28) Gray, W. R. (1967) Sequential degradation plus dansylation. *Methods Enzymol.* 11, 469.
- (29) Hartley, B., and Massey, V. (1956) The active center of chymotrypsin I. Labelling with a fluorescent dye. *Biochim. Biophys. Acta* 21, 58.
- (30) Titus, J., Haugland, R., Sharrow, S. O., and Segal, D. M. (1982) Texas Red, a hydrophilic, red-emitting fluorophore for use with fluorescein in dual parameter flow microfluorometric and fluorescence microscopic studies. *J. Immunol. Methods* 50, 193.
- (31) Wadsworth, P., and Salmon, E. (1986) Preparation and characterization of fluorescent analogs of tubulin. *Methods Enzymol.* 134, 519.
- (32) De Belder, A. N., and Granath, K. (1973) Preparation and properties of fluorescein labeled dextrans. *Carbohydr. Res.* 30, 375.
- (33) Gurd, F. R. N. (1967) Carboxymethylation. *Methods Enzymol.* 11, 532.
- (34) Shiao, D. D. F., Lumry, R., and Rajender, S. (1972) Modification of protein properties by change in charge. *Eur. J. Biochem.* 29, 377.
- (35) Singh, P. (1977) Carbamazepine antigens and antibodies. U.S. Patent 4,058,511.
- (36) Lundblad, R. L., and Noyes, C. M. (1984) *Chemical Reagents for Protein Modification*, Vol. I, p 55, CRC Press, New York.
- (37) Ansorge, W. (1988) Non-radioactive automated sequencing of oligonucleotides by chemical degradation. *Nucleic Acids Res.* 16, 2203.
- (38) Johnson, A. E., Adkins, H. J., Matthews, E. A., and Cantor, C. R. (1962) Distance moved by transfer RNA during translocation from the A site to the P site on the ribosome. *J. Mol. Biol.* 156, 113.
- (39) Smyth, D. G., Blumenfeld, O. O., and Konigsberg, W. (1964) Reaction of *N*-ethylmaleimide with peptides and amino acids. *Biochem. J.* 91, 589.

- (40) Brown, R. D., and Matthews, K. S. (1979) Chemical modification of lactose repressor proteins using N-substituted maleimides. *J. Biol. Chem.* 254, 5128.
- (41) Ishi, S. S., and Lehrer, J. (1966) Effects of the state of the succinimido-ring on the fluorescence and structural properties of pyrene maleimide labeled alpha-tropomyosin. *Biophys. J.* 50, 75.
- (42) Kosower, N. S. (1979) Bimane fluorescent labels: labeling of normal human red cells under physiological conditions. *Proc. Natl. Acad. Sci. U.S.A.* 76, 3382.
- (43) Carlsson, J., Drevin, H., and Azen, R. (1978) Protein thiolation and reversible protein-protein conjugation. N-succinimidyl 3-(2-pyridyldithio)propionate, a new heterobifunctional reagent. *Biochem. J.* 173, 723.
- (44) Jentoft, J. E., and Dearborn, P. G. (1979) Labeling of proteins by reductive methylation using sodium cyanoborohydride. *J. Biol. Chem.* 254, 4359.
- (45) Jue, R., Lambert, J. M., Pierce, L. R., and Traut, R. R. (1978) Addition of sulfhydryl groups to *Escherichia coli* ribosomes by protein modification with 2-iminothiolane (methyl 4-mercaptobutyrimidate). *Biochemistry* 17, 5399.
- (46) McCall, M. J., Diril, H., and Meares, C. F. (1990) Simplified method for conjugating macrocyclic bifunctional chelating agents to antibodies via 2-iminothiolane. *Bioconjugate Chem.* 1, 222.
- (47) Julian, R. (1983) A new reagent which may be used to introduce sulfhydryl groups into proteins, and its use in the preparation of conjugates for immunoassay. *Anal. Biochem.* 132, 68.
- (48) Ghetie, V., Till, M. A., Ghetie, M., Tucker, T., Porter, J., Patzer, E. J., Richardson, J. A., Uhr, J. W., and Vitetta, A. (1990) Preparation and characterization of conjugates of recombinant CD4 and deglycosylated ricin A chain using different cross-linkers. *Bioconjugate Chem.* 1, 24.
- (49) Cumber, J. A., Forrester, J. A., Foxwell, B. M. J., Ross, W. C. J., and Thorpe, P. E. (1985) Preparation of antibody-toxin conjugates. *Methods Enzymol.* 112, 207.
- (50) Haugland, R. P. (1989) *Handbook of Fluorescent Probes and Research Chemicals*, p 54, Molecular Probes, Inc., Eugene, OR.
- (51) Youle, R. J., and Neville, D. M. (1980) Anti-Thy 1.2 monoclonal antibody linked to ricin is a potent cell-type specific toxin. *Proc. Natl. Acad. Sci. U.S.A.* 77, 5483.
- (52) Ghosh, S. S., Kao, P. N., McCue, A. W., and Chappelle, H. L. (1990) Use of maleimide-thiol coupling chemistry for efficient synthesis of oligonucleotide-enzyme conjugate hybridization probes. *Bioconjugate Chem.* 1, 71.
- (53) (a) Komatsu, S. K., Devries, A. L., and Feeney, R. E. (1970) Studies of the structure of freezing point-depressing glycoproteins from an Antarctic fish. *J. Biol. Chem.* 245, 2909. (b) Vanderheede, J., Ahmed, A. I., and Feeney, R. E. (1972) Structure and role of carbohydrate in freezing point-depressing glycoproteins from an Antarctic fish. *J. Biol. Chem.* 247, 7885.
- (54) Freytag, J. W. (1984) Affinity column-mediated immunoenzymetric assays: influence of affinity column ligand and valency of antibody-enzyme conjugates. *Clin. Chem.* 30, 1494.
- (55) Thorpe, P. E. (1987) New coupling reagents for the synthesis of immunotoxins containing a hindered disulfide bond with improved stability *in vivo*. *Cancer Res.* 47, 5924.
- (56) Ji, L., and Ji, T. H. (1981) Both α and β subunits of human chorionic gonadotropin photoaffinity label the hormone receptor. *Proc. Natl. Acad. Sci. U.S.A.* 78, 5465.
- (57) Keana, J. F. W., and Cai, S. X. (1989) functionalized perfluorophenyl azides: New reagents for photoaffinity-labeling. *J. Fluorine Chem.* 43, 151.
- (58) Crocker, P. J., Imai, N., Rajagopalan, K., Boggess, M. A., Kwiatkowski, S., Dwyer, L. D., Vanaman, T. C., and Watt, D. S. (1990) Heterobifunctional cross-linking reagents incorporating perfluorinated aryl azides. *Bioconjugate Chem.* 1, 419.
- (59) Batra, S. P., and Nicholson, B. H. (1982) 9-Azidoacridine, a new photoaffinity label for nucleotide and aromatic binding sites in proteins. *Biochem. J.* 207, 101.
- (60) Hixson, S. H., and Hixson, S. S. (1975) p-Azidophenacyl bromide, a versatile photolabile bifunctional reagent. Reaction with glyceraldehyde-3-phosphate dehydrogenase. *Biochemistry* 14, 4251.
- (61) Bayley, H. (1983) *Photogenerated Reagents in Biochemistry and Molecular Biology*. Elsevier, New York.
- (62) Tao, T., Lamkin, M., and Scheiner, C. (1984) Studies on the proximity relationships between thin filament proteins using benzophenone-4-maleimide as a site-specific photoreactive crosslinker. *Biophys. J.* 45, 261.
- (63) Valdes-Aguilera, O., and Neckers, D. C. (1989) Aggregation phenomena in xanthene dyes. *Acc. Chem. Res.* 22, 171.
- (64) Midoux, P., Roche, A. C., and Monsigny, M. (1987) Quantitation of the binding, uptake, and degradation of fluorescently labeled neoglycoproteins by flow cytometry. *Cytometry* 8, 327.
- (65) Stryer, L., and Haugland, R. P. (1967) Energy transfer: A spectroscopic ruler. *Proc. Natl. Acad. Sci. U.S.A.* 58, 719.
- (66) Zuk, R. F., Rowley, G. L., and Ullman, E. F. (1979) Fluorescence protection immunoassay: A new homogeneous assay technique. *Clinical Chem.* 25, 1554.
- (67) Whitaker, J. E., Haugland, R. P., Moore, P. L., Hewitt, P. C., Reese, M., and Haugland, R. P. Cascade Blue derivatives: water soluble, reactive, blue emission dyes evaluated as fluorescent labels and tracers. *Anal. Biochem.* In press.
- (68) Spack, E. G., Jr., Packare, B., Wier, M. L., and Edidin, M. (1986) Hydrophobic adsorption chromatography to reduce nonspecific staining by rhodamine-labeled antibodies. *Anal. Biochem.* 158, 233.
- (69) Carraway, K. L., and Koshland, D. E., Jr. (1968) Reaction of tyrosine residues in proteins with carbodiimide reagents. *Biochem. Biophys. Acta* 160, 272.
- (70) Smyth, D. G. (1967) Acetylation of amine and tyrosine hydroxyl groups. *J. Biol. Chem.* 242, 1592.
- (71) Van Dalen, J. P. R., and Haajlman, J. J. (1974) Determination of the molar absorption coefficient of bound tetramethylrhodamine isothiocyanate relative to fluorescein isothiocyanate. *J. Immunol. Methods* 5, 103.
- (72) Wessendorf, M. W., Tallaksen-Greene, S. J., and Wohlueter, R. M. (1990) A spectrophotometric method for determination of fluorophore-to-protein ratios in conjugates of the blue fluorophore 7-amino-4-methylcoumarin-3-acetic acid (AMCA). *J. Histochem. Cytochem.* 38, 87.
- (73) Guar, R. K., and Gupta, K. C. (1989) A spectrophotometric method for the estimation of amino groups on polymer supports. *Anal. Biochem.* 180, 253.
- (74) Srivastava, P. C., Buchsbaum, D. J., Allred, J. F., Brubaker, P. G., Hanna, D. E., and Spiker, J. K. (1990) A new conjugating agent for radioiodination of proteins: Low *in vivo* deiodination of a radiolabeled antibody in a tumor model. *Biotechniques* 8, 536.
- (75) Vlgers, G. P. A., Cove, M., and McIntosh, J. R. (1988) Fluorescent microtubules break up under illumination. *J. Cell. Biol.* 107, 1011.
- (76) Kellogg, D. R., Michison, T. J., and Alberts, B. M. (1988) Behavior of microtubules and actin filaments in living *Drosophila* embryos. *Development* 103, 675.
- (77) Khalfan, H. (1986) Aminomethylcoumarin acetic acid: a new fluorescent labelling reagent for proteins. *Histochem. J.* 18, 497.
- (78) Gorbaky, G. J., Sammak, P. J., and Borisy, G. G. (1988) Microtubule dynamics and chromosome motion visualized in living anaphase cells. *J. Cell. Biol.* 106, 1185.
- (79) Hoffman, K., Finn, F., and Kiso, Y. (1978) Avidin-biotin affinity columns. General methods for attaching biotin to peptides and proteins. *J. Am. Chem. Soc.* 100, 3585.
- (80) Bolton, A. E., and Hunter, W. M. (1973) The labelling of proteins to high specific radioactivities by conjugation to a ^{125}I -containing acylating agent. Application to the radioimmunoassay. *Biochem. J.* 133, 529.
- (81) Duncan, J. S., Weston, P. D., and Wrigglesworth, R. (1982) A new reagent which may be useful to introduce sulfhydryl groups into proteins, and its use in the preparation of conjugates for immunoassay. *Anal. Biochem.* 132, 68.

Exhibit 7

Review

Protein engineering of subtilisin

Philip N. Bryan *

*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive,
Rockville, MD 20850, USA*

Received 21 March 2000; received in revised form 17 August 2000; accepted 28 September 2000

Abstract

The serine protease subtilisin is an important industrial enzyme as well as a model for understanding the enormous rate enhancements affected by enzymes. For these reasons along with the timely cloning of the gene, ease of expression and purification and availability of atomic resolution structures, subtilisin became a model system for protein engineering studies in the 1980s. Fifteen years later, mutations in well over 50% of the 275 amino acids of subtilisin have been reported in the scientific literature. Most subtilisin engineering has involved catalytic amino acids, substrate binding regions and stabilizing mutations. Stability has been the property of subtilisin which has been most amenable to enhancement, yet perhaps least understood. This review will give a brief overview of the subtilisin engineering field, critically review what has been learned about subtilisin stability from protein engineering experiments and conclude with some speculation about the prospects for future subtilisin engineering. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Folding; Stability; Site-directed mutagenesis; Design; Directed evolution

1. Overview

In March of 1985, the first UCLA Symposium on Protein Structure, Folding and Design convened in Keystone Colorado [105]. The atmosphere reflected a distinct giddiness among many of us about the prospects of the newly anointed field of 'Protein Engineering' [170]. The meeting was timely because in the early 1980s a number of technical breakthroughs came together which enabled the introduction of specific mutations into a gene, heterologous expression of the altered protein, and relatively rapid assessment of the structural consequences of the mutations by X-ray structure determination. In the keynote

address, however, Frederick Richards of Yale University asserted that while site-directed mutagenesis was fun, it was really just the next phase of chemical modification and unlikely to revolutionize understanding of protein folding and enzymology. After 15 years and thousands of site-directed mutants, it probably can be said that a good time has been had by all. But given the perspectives of time and experience, what has been accomplished from protein engineering? This review will give a brief overview of the subtilisin engineering field, critically review what has been learned about subtilisin stability from protein engineering experiments and conclude with some speculation about the prospects for future subtilisin engineering.

Mutations in well over 50% of the 275 amino acids of subtilisin have been reported in the scientific literature (Table 1). Many more examples exist in the

* Fax: +1 301 738 6255; E-mail: bryan@umbi.umd.edu

patent literature and undoubtedly still more lurk unfathomed in the freezers of biotechnology companies. Subtilisins constitute a large class of microbial, serine proteases, but the ones most mutagenized are those secreted from the *Bacillus* species *amyloliquefaciens* (BPN'), *subtilis* (subtilisin E) and *lentus* (savinase). Subtilisins are important industrial enzymes as well as models for understanding the enormous rate enhancements affected by enzymes. For these reasons, along with the timely cloning of the gene, ease of expression and purification and availability of atomic resolution structures, subtilisin became a model system for protein engineering studies.

Protein engineering of subtilisin commenced in the mid 1960s when the active site serine 221 was converted to cysteine through chemical modification [101,119]. As it turned out, this first alteration remains one of the most useful. C221 subtilisin is catalytically wounded to the point that it will barely hydrolyze peptide bonds but turns out to be quite reactive with certain activated ester substrates [115,116]. This combination of properties has made it a useful tool for catalyzing synthetic reactions. These include condensation of amino acids to form peptides and transesterification reactions such as regioselective acylation of sugars [83,98,187,188].

The first genetic modifications in subtilisin occurred rapidly after the gene was cloned in the early 1980s [72,171,182]. The early standard for genetic manipulation was subtilisin BPN', which was engineered for stability [26,47,183], catalytic mechanism [20,168,180] and substrate specificity [46]. The rationales for modifying subtilisin have expanded over the years to include the following eight broad classifications:

(1) Catalytic mechanism: [15,20,31,32,36,41,97,101,102,104,119–121,129,130,147,148,168,169,178,180,185].

(2) Substrate specificity: [5,6,8,9,28–30,38–40,46,56–58,85,89–91,94,122,123,144,155,156,161,163–165,167,179,181,184].

(3) New activities: [1,3,10,11,60–63,79,114,117,134,152,193].

(4) General proteolytic activity: [54,77,153,154,157,159].

(5) General stability: [4,16,22,23,25–27,34,35,45,48,53,65,74,75,78,80,95,96,99,100,107,110–112,124,132,145,158,160,166,183,190,191,194].

(6) Stability in exotic environments: [33,47,55,109,149,174,186].

(7) Surface activity: [17,18,44,69].

(8) Folding mechanisms: [19,21,24,42,43,49–51,67,68,73,82,86–88,127,128,131,133,138–142,150,151,172,176,177].

Most subtilisin engineering continues to involve catalytic amino acids, substrate binding regions and stabilizing mutations. Included in the active site category are mutations of the catalytic triad (D32, H64, S221), the oxyanion hole (N155) and mutations which influence pK_a of H64 through long range electrostatics. Most mutations affecting specificity have been made in the binding pockets S1 and S4 [12]. The S1 amino acids comprise positions 127, 152, 154, 156 and 166 and the S4 amino acids comprise positions 102, 104, 107, 126 and 128. A excellent review of the use of protein engineering to understand catalytic mechanism and substrate specificity appeared in 1995 [113].

2. Subtilisin stability

Stability has been the property of subtilisin which has been most amenable to enhancement, yet perhaps least understood. Rationalizing stability increases resulting from mutation in structural and energetic detail is limited by the inability to study the folding reaction under equilibrium conditions. The most basic protein stability experiment is determining the free energy of unfolding [70,162]. This question is still not resolved for subtilisin. Biosynthesis of subtilisin requires participation of an N-terminal prodomain [71]. The folding rate of mature subtilisin without the prodomain occurs on a time scale closer to geological than biological. By combining biochemical analysis with information from mutagenesis experiments, however, one can now make an informed estimate of the free energy of folding mature subtilisin and use this information to better evaluate stabilizing mutations.

2.1. Energetics of the subtilisin folding reaction

2.1.1. Calcium binding

A fundamental variable to address in subtilisin stability is its colossal calcium dependence [52,175].

Table I

No.	BP
1	A
2	Q
3	S
4	V
5	P
6	Y
7	G
8	V
9	S
10	Q
11	I
12	K
13	A
14	P
15	A
16	L
17	H
18	S
19	Q
20	G
21	Y
22	T
23	G
24	S
25	N
26	V
27	K
28	V
29	A
30	V
31	I
32	D
33	S
34	G
35	I
36	D
37	S
38	S
39	H
40	P
41	D
42	L
43	K
44	V
45	A
46	G
47	G
48	A
49	S
50	M
51	V

Table 1

No.	BPN'	Mutation
1	A	C w/C78 [108]
2	Q	R [23]; E, K, R, L, W [149]
3	S	C w/C206/Δ75–83 [149]; T [3]
4	V	I [53]
5	P	A, S w/Δ75–83 [149]
6	Y	
7	G	
8	V	I [25]
9	S	F [191]
10	Q	
11	I	
12	K	
13	A	
14	P	L [191]
15	A	K
16	L	
17	H	
18	S	
19	Q	E [45]
20	G	
21	Y	
22	T	C w/C87 [110,183]
23	G	
24	S	C w C87 [110,183]
25	N	
26	V	C w/235 [108]; C w/232 [95]; R [45]
27	K	C w/C89 [108]; R [54,65]
28	V	
29	A	C w/C119 [95]
30	V	
31	I	L [157]
32	D	N, A [31]; N [51]
33	S	D, E [5]
34	G	
35	I	
36	D	Q [148]; C w/C210 [95]; insertion of D (savinase) [174] [191]
37	S	
38	S	
39	H	
40	P	
41	D	C w/C80 [95]; Q, A w/Δ75–83 [149]
42	L	
43	K	N [134]; N, R, w/Δ75–83 [149]
44	V	
45	A	Replacement 45–63 with thermitase sequence [16]
46	G	
47	G	
48	A	
49	S	D, R [75]; P [65]
50	M	F [35,111]
51	V	K [45]

Table 1 (continued)

No.	BPN'	Mutation
52	P	
53	S	T [124]
54	E	
55	T	
56	N	
57	P	
58	F	
59	Q	R
60	D	N (subt E) [33,194]
61	N	C w/C98 [160]
62	N	D [5]; CMM [36]
63	S	D [25]
64	H	A [31]
65	G	
66	T	
67	H	Y, A [3]
68	V	C [7]
69	A	
70	G	A, S w/Δ75–83 [149]
71	T	V [53]
72	V	I [153]
73	A	L, H w/Δ75–83 [149]
74	A	
75	L	Deletion 75–83 [19]
76	N	D [99,111,174,191]
77	N	D [45]
78	S	C w/C1 [108]; D [25]
79	I	T
80	G	C w/C41 [95]
81	V	
82	L	
83	G	
84	V	
85	A	C w/232 [108]
86	P	
87	S	C w/22 and 24 [110,183]; S (savinase) [54]
88	A	
89	S	E [45]; E89S (savinase) [65]
90	L	
91	Y	
92	A	T [153]
93	V	I [190]
94	K	
95	V	
96	L	
97	G	D97G (subt E) [33]
98	A	K [45]; C w/C61 [160]
99	D	S, K [147]
100	G	A, V, L [164]
101	S	H, K, E [165]
102	G	F [9]
103	Q	R [33,194]; A [54]

Table 1 (continued)

104	Y	A, R, D, F, S, W, Y [8]; W [167]; A, F [122,123]; V [174]; D [6]; I [54]
105	S	
106	W	
107	I	V [35]; G, A, V, L, F [144]; G, A, V [123]
108	I	
109	N	S [99,190]
110	G	
111	I	
112	E	
113	W	
114	A	
115	I	
116	A	T, E [124]
117	N	
118	N	S [34,191,194]
119	M	C w/C29 [95]
120	D	H120D (savinase) [174]
121	V	
122	I	C w/C 147 [108]
123	N	S [54]
124	M	L, I [3]
125	S	A, G [3]
126	L	I [124]; A, F [144]; G, A, V [123]
127	G	A, S, V [156]
128	G	F [9]; S128G (savinase) [174]
129	P	
130	S	F [9]
131	G	D [33,124,166]; H, K [165]
132	S	F [9]
133	A	
134	A	
135	L	A, V, F [144]
136	K	
137	A	
138	A	
139	V	
140	D	
141	K	
142	A	
143	V	
144	A	
145	S	
146	G	
147	V	C w/C122 [108]
148	V	C w/243 [95]
149	V	
150	V	
151	A	
152	A	C, S [3]
153	A	
154	G	A, R, L, F, P, T [161]
155	N	L [20]; A, L, H, Q, R [180]

Table 1 (continued)

156	E	Q, S [184]; S, K [147]; G [33]; CMM [36]
157	G	
158	T	158–165 replacement with thermitase sequence [16]
159	S	
160	G	
161	S	C [191]; deletion 161–164 [155]
162	S	
163	S	
164	T	R [45]; S164T (savinase) T [53]
165	V	C w/C191 [108]
166	G	A, S, C, T, P, V, L, I, F, Y, W [46]; D, E, Q, M, K, R [184]; S [124]; D [5]; R [191]; CMM [36]
167	Y	
168	P	
169	G	A [111,181]
170	K	Y, L, M [65]
171	Y	
172	P	D, E [112]
173	S	
174	V	
175	I	
176	A	
177	V	
178	G	
179	A	
180	V	
181	D	S [33]; N [134]; D [190]
182	S	G [33]
183	S	
184	N	
185	Q	
186	R	
187	A	
188	S	P [33,124,132]
189	F	
190	S	
191	S	C w/C165 [108]
192	V	T [191]
193	G	
194	P	S194P (subt E) [65,191]; A194P (savinase) [174]
195	E	Q [74]; Q, E, D, F, M, K, R (savinase) [65,174]
196	L	
197	D	N [65,166]
198	V	
199	M	
200	A	
201	P	
202	G	
203	V	K [17]

Table 1 (continued)

204	S
205	I
206	Q
207	S
208	T
209	L
210	P
211	G
212	N
213	K
214	Y
215	G
216	A
217	Y
218	N
219	G
220	T
221	S
222	M
223	A
224	S
225	P
226	H
227	V
228	A
229	G
230	A
231	A
232	A
233	L
234	I
235	L
236	S
237	K
238	H
239	P
240	N
241	W
242	T
243	N
244	T
245	Q
246	V
247	R
248	S
249	S
250	L
251	E
252	N
253	T
254	T

Table 1 (continued)

204	S	F [17]
205	I	V205I (savinase) [53]
206	Q	C [111]; C w/C3/Δ75–83 [149]; N, D, Y, E, K, I, F, L, W [17]
207	S	
208	T	
209	L	F [17]
210	P	C w/C36 [95]
211	G	K, P, L, W [96]
212	N	P, A, V, S [96]
213	K	R [35]; T [147]
214	Y	K w/Δ75–83 [149]
215	G	
216	A	E [17]
217	Y	L [181]; K [111]; W [134]; CMM [36]
218	N	S, T, A, C, D, W [26]; S [99,111,190]; M [17]; S, T, A, H [3]
219	G	
220	T	A [15]
221	S	C [1,101,119]; A [31]; seleno [10]
222	M	All [47]; Me-S-C [55]; A [134,194]; G, S, A, V, F [3]
223	A	S [3]
224	S	A, C [3]
225	P	A [1]; G [3]
226	H	
227	V	
228	A	
229	G	
230	A	
231	A	
232	A	C w/C85 [108]; C w/C26 [95]
233	L	
234	I	
235	L	R [45]; K235L (savinase) [174]
236	S	
237	K	
238	H	
239	P	G, K, R [158]
240	N	
241	W	
242	T	
243	N	C w/C148 [95]
244	T	
245	Q	
246	V	
247	R	
248	S	N, A, L [66]
249	S	C w/C273 [108]
250	L	
251	E	E [65]
252	N	
253	T	C w/C273 [108]
254	T	A [124]

Table 1 (continued)

255	T	A [33]
256	K	Y [134]
257	L	
258	G	
259	D	
260	S	
261	F	
262	Y	
263	Y	
264	G	
265	K	
266	G	
267	L	
268	I	
269	N	D
270	V	
271	Q	E [2,45]; G [65]
272	A	
273	A	C w/C249 or C253 [108]
274	A	A (savinase) [54]
275	Q	

A universal feature of subtilisins is the presence of one or more calcium binding sites. High resolution X-ray structures of subtilisin BPN', as well as several homologues [13,14,59,93], have revealed details of a conserved, calcium binding site, termed site A. Calcium at site A is coordinated by five carbonyl oxygen ligands and one aspartic acid. Four of the carbonyl oxygen ligands to the calcium are provided by a loop comprising amino acids 75–83. The geometry of the ligands is that of a pentagonal bipyramid whose axis runs through the carbonyls of 75 and 79. On one side of the loop is the bidentate carboxylate (D41), while on the other side is the N-terminus of the protein and the side chain of Q2. The seven coordination distances range from 2.3 to 2.6 Å, the shortest being to the aspartyl carboxylate. Three hydrogen bonds link the N-terminal segment to loop residues 78–82 in parallel-β arrangement.

Because of the marginal stability of subtilisin without calcium bound, the energetics of calcium binding at site A are difficult to study independently of the unfolding reaction. By employing an inactive and stabilized version of subtilisin, the calcium-free (apo) form of subtilisin can be produced and calcium binding measured by microcalorimetry and fluorescence spectroscopy [19]. The binding parameters obtained by titration calorimetry are

$\Delta H = -11$ kcal/mol and $K_a = 7 \times 10^6$ M⁻¹ at 25°C. The standard free energy of binding is 9.3 kcal/mol, so the binding of calcium is primarily enthalpically driven with only a small net loss in entropy ($\Delta S_{\text{binding}} = -6.7$ cal/°mol). This is surprising since transfer of calcium into water results in a loss of entropy of -60 cal/°mol. Therefore the freeing of water upon calcium binding to the protein will make a major contribution to the overall ΔS of the process. The gain in solvent entropy upon binding must be compensated for by a loss in entropy of the protein. Presumably, the loop amino acids 75–83 and the first few N-terminal residues have increased mobility when calcium is absent from the A site.

A second ion binding site (site B) is located 32 Å from site A in a shallow crevice between two segments of polypeptide chain near the surface of the molecule. The coordination geometry of this site closely resembles a distorted pentagonal bipyramid. Three of the formal ligands are derived from the protein and include the carbonyl oxygen atom of E195 and the two side chain carboxylate oxygens of D197. Four water molecules complete the first coordination sphere. Evidence that site B binds calcium comes from determining the occupancy of the site in a series of X-ray structures from crystals grown in 50 mM NaCl with calcium concentrations ranging from 1 to 40 mM [112]. In the absence of excess calcium, this locus was found to bind a sodium ion. The binding of these two ions appears to be mutually exclusive so that as the calcium concentration increases, the sodium ion is displaced, and a water molecule appears in its place directly coordinated to the bound calcium [112]. Analysis of occupancy vs. calcium concentration indicates that K_a is approx. 40 M⁻¹.

2.1.2. Calcium-independent stability

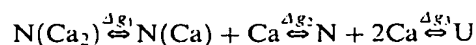
Subtilisin does not refold to the native state on an observable time scale except under conditions which make direct measurements of the equilibrium constant for folding impractical [64]. Site-directed mutagenesis afforded an opportunity to simplify the subtilisin folding reaction and test whether a calcium-free mutant subtilisin might fold more readily than the wild type protein. The calcium binding loop is formed from a nine amino acid bubble in the last

turn of a 14-residue α -helix involving amino acids 63–85 [93]. Deleting amino acids 75–83 creates an uninterrupted helix and abolishes the calcium binding potential at site A [2,19]. The X-ray structure has shown that except for the region of the deleted calcium binding loop, the structure of the mutant and wild type protein are remarkably similar considering the size of the deletion. The structures of subtilisin with and without the deletion superimpose with an rms difference between 261 C α positions of 0.17 Å. The N-terminus of the wild type protein lies beside the site A loop, furnishing one calcium coordination ligand, the side chain oxygen of Q2. In $\Delta 75$ –83 subtilisin, the loop is gone, leaving residues 1–4 disordered, but the helix is uninterrupted and shows normal helical geometry over its entire length.

The folding rate of $\Delta 75$ –83 BPN' is much faster than BPN'. Although it is hard to compare their folding rates under similar conditions [64,92], it is certain that $\Delta 75$ –83 BPN' folds at least 10⁴ times faster than BPN' in 0.1 M KPi, pH 7.0. The unfolding rates of the apo form of BPN' and $\Delta 75$ –83 BPN' are very similar [19]. Since $\Delta G_{\text{unfolding}} = -RT \ln(k_{\text{unfolding}}/k_{\text{folding}})$ in a two state system, the simplest interpretation of the unfolding and refolding rates would mean that $\Delta G_{\text{unfolding}}$ for $\Delta 75$ –83 BPN' is at least 5.5 kcal/mol greater at 25°C than for apo BPN'. Recent H-D exchange data indicate that the total $\Delta G_{\text{unfolding}}$ for $\Delta 75$ –83 BPN' is approx. 7 kcal/mol in 0.1 M KPi, pH 7.0 and 25°C (unpublished data). This would mean that apo BPN' is near the margin of thermodynamic stability.

2.1.3. Calcium-dependent stability

In view of the marginal stability of apo-subtilisin, it is evident that calcium binding makes a dominant contribution to conformational stability. By binding at a specific site in the tertiary structure, calcium contributes its binding energy to the stability of the native state and contributes to the overall free energy of folding. The unfolding reaction of subtilisin BPN' can be divided as follows:



where $N(\text{Ca}_2)$ is the native form of subtilisin with calcium bound to both sites, $N(\text{Ca})$ is the native form of subtilisin with calcium bound to site A, N

is the folded tein. The te equal to ΔG one can cal free energy

$$\Delta G_{\text{binding}} =$$

Thus the c subtilisin in 25°C. The c in 10 mM kcal/mol. Tl which concl responsible rate of sub concentrations examination light of a b subtilisin fol effect on su erate concer

2.2. Kinetic

In most stability is c a function inactivation todigestion, tain amino ity by this inactivation determined. vated with of the meth. If this occur enzyme rem autodigestio tant mecha tions of enzy In general, of inactivati measuring t becomes the activation a seen by dire subtilisin E with the rat

is the folded apoprotein and U is the unfolded protein. The total free energy of unfolding is therefore equal to $\Delta G_1 + \Delta G_2 + \Delta G_3$. From the binding constant, one can calculate the contribution of calcium to the free energy of subtilisin folding from the equation:

$$\Delta G_{\text{binding}} = -RT \ln(1 + K_a[\text{Ca}])$$

Thus the contribution of site A to the stability of subtilisin in 10 mM calcium is 6.6 kcal/mol at 25°C. The contribution of calcium binding to site B in 10 mM calcium and 50 mM sodium is only 0.2 kcal/mol. This analysis is at odds with earlier studies which concluded that calcium binding to site B is responsible for the large decrease in the inactivation rate of subtilisin in the presence of millimolar concentrations of calcium [16,112]. As shown below, re-examination of calcium-dependent stability data in light of a better understanding of the energetics of subtilisin folding shows that site B has relatively little effect on subtilisin stability in the presence of moderate concentrations of monovalent cations.

2.2. Kinetics of irreversible inactivation

In most protein engineering studies of subtilisin, stability is defined in terms of the loss of activity as a function of time. The mechanisms of irreversible inactivation can be complex involving unfolding, autolysis, aggregation and chemical damage to certain amino acids. If one wishes to understand stability by this definition, the rate determining step in inactivation under the specified condition must be determined. For example, subtilisin can be inactivated with hydrogen peroxide due to the oxidation of the methionine next to the active site serine [146]. If this occurs, it is irrelevant to activity whether the enzyme remains folded or not. It is also clear that autolysis will become a relatively more important mechanism of inactivation at high concentrations of enzyme because it is a second order reaction. In general, however, studies which measure the rate of inactivation at elevated temperature are indirectly measuring the rate of unfolding because unfolding becomes the rate determining step in irreversible inactivation as temperature is increased. This can be seen by directly comparing the rate of unfolding of subtilisin BPN' using calorimetric measurements with the rate of inactivation under the same condi-

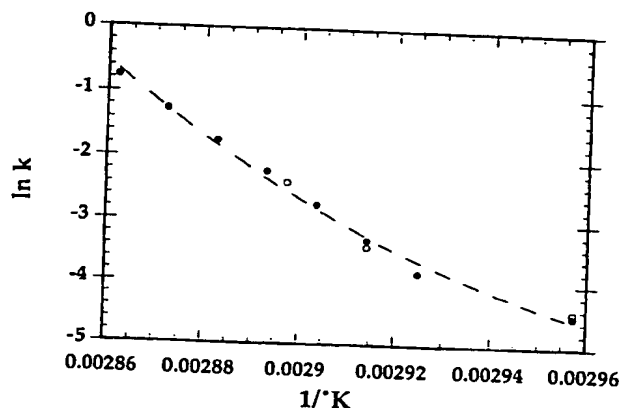


Fig. 1. Comparison of the rates of irreversible thermal inactivation of subtilisin BPN' with the rate of thermal unfolding in 50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 10 mM CaCl₂, over the temperature range of 65–75°C. Unfolding rates are measured by differential scanning calorimetry. Data are plotted as the natural logarithm of the rate constants vs. 1/K. Solid circles show the rate of unfolding and open circles show the rate of inactivation. The activation energy of both processes is approx. 80 kcal/mol at 65°C.

tions (Fig. 1). Hence changes in rate of irreversible inactivation at elevated temperatures resulting from mutation are reflecting a change in activation energy for unfolding.

Stabilizing mutations in subtilisin characterized by changes in the kinetics of inactivation can be classified into three groups: (1) stabilizing only in calcium, (2) stabilizing only in chelants, and (3) stabilizing in both conditions (Table 2). From this partitioning it is evident that the mechanism of thermal inactivation differs depending on whether the calcium sites are occupied. To understand why this is so, one must understand how the kinetics of inactivation are related to the kinetics of unfolding and how the kinetics of unfolding are related to the kinetics of calcium loss.

2.2.1. Inactivation in excess EDTA

Thermal inactivation in EDTA is a two step process as shown in mechanism 1:

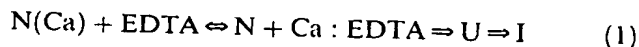


Fig. 2 compares the rate of calcium dissociation with the rate of unfolding as a function of temperature for an inactive variant of subtilisin BPN' [19]. Repartitioning of calcium from site A into a strong chelator

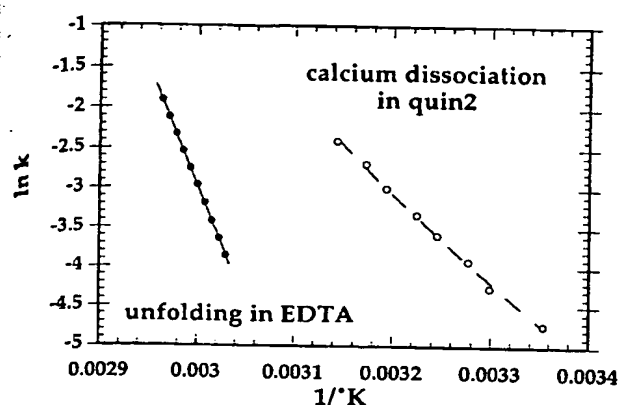
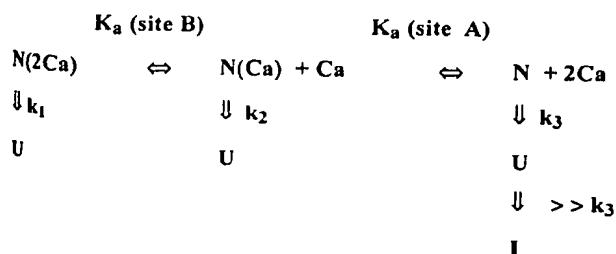


Fig. 2. Comparison of the rates of calcium dissociation in excess fluorescent chelator (quin2) with the rate of thermal unfolding, for the inactive subtilisin mutant, S11 [19]. The activation energies are 23 kcal/mol for calcium dissociation in quin2 and 60 kcal/mol for unfolding in 50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 10 mM EDTA, at 45°C. Data are plotted as the natural logarithm of the rate constants vs. $1/T$. Solid circles show the rate of unfolding and closed circles show the rate of calcium dissociation.

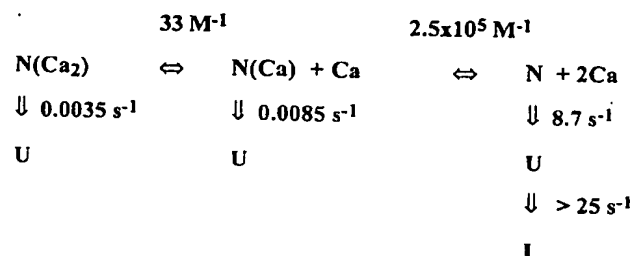
calcium at site A ($K_{S-Ca} = 7 \times 10^6 \text{ M}^{-1}$) and the binding constant of EDTA for calcium ($K_{E-Ca} = 2 \times 10^8 \text{ M}^{-1}$), then less than 0.02% subtilisin would be bound to calcium at equilibrium. Examples of mutations which stabilize apo-subtilisin are M50F and the disulfides C22-C87 and C206-C216. The irony is that a mutation which preferentially stabilizes apo-subtilisin relative to the bound form, will weaken calcium binding and catalyze inactivation under conditions of excess calcium and high temperature (see mechanism 2 below). This phenomenon is displayed in the M50F mutant, which is more stable than wild type in 10 mM EDTA but less stable in 10 mM CaCl_2 (Table 2).

2.2.2. Inactivation in excess calcium

The inactivation of subtilisin in excess calcium is diagrammed in mechanism 2:



In excess calcium (e.g. $\geq 1 \text{ mM}$) and moderate temperature, calcium binding and dissociation is in rapid equilibrium because calcium binding is much faster than unfolding. The rate of inactivation is determined by the fraction of each native species times its unfolding rate. Using mechanism 2, one can show that calcium dependent stabilization of subtilisin is dominated by site A rather than site B. Fig. 3 plots the rate of inactivation of BPN' at 65°C as a function of calcium concentration and fits the data to the following mechanism:



The mechanism predicts that K_a values of site A and site B are $2.5 \times 10^5 \text{ M}^{-1}$ and 33 M^{-1} at 65°C. The rate of inactivation of subtilisin with only site A occupied ($N(Ca)$) is about 1000 times slower than apo-subtilisin (N) and the rate of inactivation with both sites occupied ($N(Ca_2)$) is about 2.5 times slower than with only site A occupied. The second prediction has been borne out by measuring the calcium dependent stability of a mutant which has site B but lacks site A [149]. The rate of inactivation of this mutant is only 2.4 times slower in 10 mM CaCl_2 , 50 mM NaCl than in 10 mM EDTA, 50 mM NaCl.

Another prediction of mechanism 2 is that any mutations which stabilize only in the presence of calcium will increase the binding constant for calcium to one or both of the calcium sites. This can be either through effects on the binding sites themselves, as proposed for mutations A116E, G131D, P172D, S63D, N76D, S78D and K256Y and the thermitase loop 45–63 in BPN', or through indirect effects on conformational stability as seen for mutations V8I, S53T, L126I, G166S, G169A and T254A (Table 2). The indirect effect on calcium binding arises because apo-subtilisin displays a loss of cooperativity in the unfolding reaction [19]. Thus many mutations which stabilize in the presence of calcium do not stabilize in the presence of EDTA, because they do not influence the rate determining step in the unfolding of

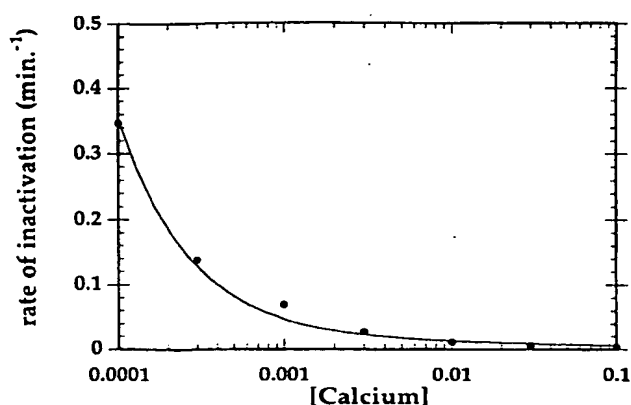


Fig. 3. The rates of thermal inactivation of subtilisin BPN' at 65°C are plotted as a function of calcium concentration. The data are fit to mechanism 3 in the text. Data taken from Fig. 1 of Pantoliano et al. [112].

apo-subtilisin. In fact, most mutations identified by random mutagenesis stabilize only in the presence of calcium. These mutants increase calcium binding affinity because they preferentially stabilize NCa relative to N. The premise that the effects of this class of mutations indirectly increase calcium affinity by increasing general stability was tested by introducing G166S, G169A and T254A into the rehabilitated S88 version of $\Delta 75-83$ subtilisin [126]. Because the unfolding of the S88 subtilisin is cooperative in EDTA, these mutations now stabilize subtilisin S88 in 50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 10 mM EDTA to approximately the same extent that they stabilize subtilisin BPN' in 50 mM Tris-HCl, pH 8.0, 50 mM NaCl, 10 mM CaCl_2 .

Finally mutations which stabilize in excess calcium and in EDTA to the same extent must stabilize N and NCa to equal extents. This would result in no change in calcium affinity. Mutations of this class are N218S, Y217K, Q206Cox and Q271E [2,111].

2.2.3. Disulfide mutants

Because of the slow rate of the subtilisin folding reaction, most stability experiments are affected only by the activation energy for unfolding and not the equilibrium constant for unfolding. This immediately explains why engineering disulfide bonds into subtilisin was so spectacularly unsuccessful in increasing resistance to thermal inactivation [95,108]. A well-

designed disulfide cross-link should stabilize a protein by decreasing the entropic cost of folding. The loss of conformational entropy in a polymer due to a cross-link has been estimated by calculating the probability that the ends of a polymer will simultaneously occur in the same volume element (v_s) according to the equation:

$$\Delta S = -R \ln(3/(2\pi l^2 N)^{3/2}) v_s$$

where N is the number of segments and l is the length of a segment [118]. Good agreement with experimental data for protein cross-linking has been achieved using $l=3.8 \text{ \AA}$ and $v_s=58 \text{ \AA}^3$, judged to be the closest approach of two -SH groups [106].

Of 18 different disulfide cross-links which have been engineered into subtilisin, three have increased stability [108,110,160]. Two of these stabilize only in the presence of EDTA. This is not surprising in retrospect because effects on the stability of the unfolded state would not generally be manifested in the activation energy of the unfolding reaction. This is because the transition state for the unfolding reaction appears to be compact, with a slightly larger heat capacity than the native state. Further analysis of one of the disulfide mutants (C22-C87) in the background of $\Delta 75-83$ BPN' showed that disulfide did in fact have the predicted effects on the unfolded state [150]. The increase in the energy of the unfolded state due to cross-linking 57 amino acids (22–87 minus the nine amino acid deletion) would be 4.2 kcal/mol at 25°C so the predicted maximum increase in folding rate at 25°C would be approx. 1000-fold. Since the 22–87 disulfide accelerated folding by 850-fold at 25°C in 0.1 M KPO_4 , pH 7.2, the acceleration of the folding rate is qualitatively consistent with the simple statistical mechanical model and suggests that amino acids 22 and 87 are ordered in the transition state for folding. Accordingly, the small influence of the disulfide on the transition state for unfolding wild type BPN' in EDTA (Table 2) indicates residues 22 and 87 are only slightly less ordered in the transition state for unfolding in EDTA than in the folded state. Other mutations which preferentially decrease the entropy of the unfolded state relative to the folded state, such as substituting for glycine or substituting with proline, also are not necessarily expected to influence the rate of unfolding.

Two
sulted in
ing. On
subtilisin
occurring
aquaticu
by rand
cross-link
link in s
fold. Th
inactivat
N-termin
 β -hairpin
between
the tran
The 3-2
 $\Delta 75-83$
ordering
state for

2.2.4. R

Rando
an effecti
even wit
of the su
jor reaso
fairly cor
bust, on
changes i
inactivati
vidual st
latively.
achieved
tein struc
tions.

Rando
ous ways
base anal
oligonucl
the abilit
increased
carried o
which all
1000 mut
stable mu
elevated
vate the v
lytic activ

Two engineered disulfide bond mutants have resulted in significant decreases in the rate of unfolding. One is a disulfide between residues 61 and 98 in subtilisin E, which was modeled after a naturally occurring disulfide in aqualysin I from *Thermus aquaticus* [160]. The other is a disulfide identified by random mutagenesis of $\Delta 75$ –83 subtilisin, which cross-links residues 3 and 206 [149]. The 61–98 cross-link in subtilisin E slows thermal inactivation by 2.3-fold. The 3–206 cross-link in $\Delta 75$ –83 subtilisin slows inactivation by 17-fold. The 3–206 disulfide links the N-terminal strand of subtilisin with the 202–219 β -hairpin. Evidently disruption of the interactions between these two structural elements is involved in the transition state for unfolding $\Delta 75$ –83 subtilisin. The 3–206 cross-link increases the folding rate of $\Delta 75$ –83 subtilisin by only 1.8-fold [126]. Evidently ordering of these residues occurs after the transition state for the folding reaction.

2.2.4. Random mutagenesis

Random mutagenesis and screening proved to be an effective method to dramatically increase stability even without much understanding of the energetics of the subtilisin folding reaction. There are two major reasons for this. First, stabilizing mutations are fairly common. Although subtilisins are naturally robust, on the order of 1% of the random amino acid changes measurably increase the half-time of thermal inactivation [124]. Second, contributions from individual stabilizing mutations generally accrue cumulatively. Thus large increases in stability can be achieved with no radical changes in the tertiary protein structure but rather minor, independent alterations.

Random mutations have been introduced in various ways, including chemical mutagens, mutagenic base analogs, error prone PCR and spiked synthetic oligonucleotides. The key element in the process is the ability to screen large numbers of mutants for increased stability. Phenotypic screening has been carried out using plate or microtiter dish assays which allows assaying proteases from approx. 100–1000 mutant clones per plate or dish. To screen for stable mutants, secreted subtilisins are incubated at elevated temperature long enough to largely inactivate the wild type enzyme. When an assay for hydrolytic activity is subsequently performed, only mutants

with stability greater than wild type will exhibit measurable activity. Once stable mutants are identified, the corresponding colony can be grown up to identify the mutation. The labor factor in screening limits the number of mutants which can be examined to the 10^4 – 10^5 range. All single amino acid substitutions in subtilisin would yield a total of 5500 different variations. Since all combinations of double substitutions would produce 3×10^7 variations, only the population of single mutations in subtilisin has been adequately searched for stabilizing events. In fact, even the population of single substitutions has not been completely explored because the nature of the genetic code dictates that each amino acid can be changed to an average of six other amino acids by a single base substitution in the gene. Thus only about 30% of the possible single substitution mutants would be produced from single base substitutions.

Early studies with chemical mutagens found eight stabilizing mutations in BPN' by screening at most 1200 different single amino acid substitutions [26,27,124]. Misincorporation induced by α -thio-deoxynucleotides identified three additional stabilizing mutations in BPN' [35] and studies using error-prone PCR to introduce mutations in subtilisin E identified 11 stabilizing mutations [191]. Five of the mutations in subtilisin E were previously identified as stabilizing in BPN'. The fact that several of the same mutations have been independently selected indicates that many of the stabilizing mutations which can be produced with single base substitutions have been identified. Since this represents only 30% of the total possible single amino acid substitutions, many other stabilizing single substitutions must exist. Two examples are the directed mutations Y217K and Q206C which both stabilize significantly but are not accessible by a single point mutation [111]. Further Miyazaki and Arnold have shown that targeting random mutagenesis to positions at which stabilizing changes were already found can identify even better amino acids at these positions [96].

Once stabilizing single amino acids changes have been identified, building a highly stable subtilisin can be accomplished in a step by step manner by combining individual mutations into the same molecule. A combination of six stabilizing changes in BPN' decreased the rate of thermal inactivation by > 300 -fold [111]. A similar result was achieved in

subtilisin E by performing multiple rounds of random mutagenesis screening and molecular breeding screening [191]. A hyperstable calcium-free subtilisin has also been constructed by a combination of design and random mutagenesis. This mutant inactivates 250 000 times more slowly than wild type BPN' in 10 mM EDTA [126,149].

3. Future prospects

3.1. Design vs. screening

What strategies will prove most effective for engineering other properties of subtilisin? At the moment directed evolution seems to have become more fashionable than structure-based design as a method to 'engineer' subtilisin. Part of this trend may be a result of earlier disappointments with the ability to predict the phenotype of designed mutants, but most is a result of advances in random mutagenesis methods [76,135,190,192]. For example, synthesis of oligonucleotides using preformed trinucleotide phosphoramidites will circumvent some of the limitations inherent to the genetic code [81]. Furthermore new methods of DNA shuffling allow efficient creation of chimeric proteases to try and combine desirable properties from parent enzymes [103,137,173]. Directed evolution and molecular breeding methods have proven useful for finding mutations which are better than wild type for several different properties [136]. There is always the danger, however, that the good will become the enemy of the best [125]. The new techniques do not circumvent the combinatorial problems inherent to purely random methods. Thus random approaches will be good for improving a global property such as stability which can be accrued incrementally but will not be successful when significant improvements depend on synergistic mutational events. Relying on the accumulation of single mutants insures that only solutions very close to the starting structure will be found. The best solutions may lie unmined a few layers deeper in mutational space.

Optimizing subtilisin activity for a specific protein sequence or for a new substrate are cases in which synergistic mutations probably will be required. Con-

sider the basic organization of the substrate binding pockets of subtilisin. Although the deep S1 and S4 binding clefts are the primary determinants of substrate specificity, subtilisin is relatively non-specific in its cleavage preferences for protein substrates. The broad specificity is in part a consequence of the fact that the substrate peptide backbone inserts itself between residues 100–104 and 125–129 to become the central strand in an antiparallel β -sheet. This is different from the more specific chymotrypsin family of proteases in which a structural equivalent of residues 100–104 is absent [113]. The best solutions to accommodate new substrates may involve altering main chain interactions and this will involve multiple synergistic mutations. When high resolution structural information becomes available for the subtilisin class of prohormone converting enzymes, it will be interesting to see what structural differences account for sequence-specific processing activity.

Introducing the bias of intelligent design into random mutagenesis experiments has been criticized because of limitations in the intelligence of designers. The dilemma is as follows. The more target positions for mutagenesis are restricted, the greater the ability of screening to identify synergistic mutations. But the greater restriction of the target positions, the greater the danger of flawed design. In many cases, however, only minimal design is required to identify productive regions of sequence space. Past experiences with directed mutagenesis have shown that mutations which have the greatest influence on substrate specificity involve either direct contacts with the substrate or electrostatic changes in the vicinity of the active site. This is also borne out by experiences with random mutagenesis and screening. For example, You, Chen and Arnold have randomly mutated subtilisin E using error-prone PCR and screened for increased activity in dimethylformamide against a defined peptide substrate [33,189]. Twelve mutations were identified in the screen. Of the twelve, two are involved in direct binding with the peptide, three are mutations of Asp or Glu to neutral amino acids at positions which would influence the pK_a of H64, five are mutations which increase general stability and only two are at positions whose connections with activity in DMF are difficult to rationalize.

3.2. Ph

Rece.
face of
sibilitie
less dire
ods for
number
four or
baries
ing all
position
that sel
selection
random
into S2
ligation.
the liga
proved
A secon
subtilisin
of the st
Selection
ried out
inhibitor

3.3. Unc

A m
method
potentia
is linked
Hence th
enzymes
phenotyp
sequence
tually m
the proc
sequence
the natu
zymes si
main ref
desired
strate, h
thesis of
activity.
Δ75–83 v
ing with

3.2. Phage display selection

Recent successes in displaying subtilisin on the surface of phagemid particles greatly expands the possibilities for selecting new properties [3,37,84]. While less direct than culture dish or microtiter plate methods for screening, phage display methods increase the number of mutants which can be screened by at least four orders of magnitude. The ability to display libraries of 1×10^9 independent mutants allows screening all combinations of amino acids at six specified positions. The obvious limitation of phage display is that selection is achieved by binding activity, so that selection of a catalytic event is not trivial. In one case random mutations at 25 positions were introduced into S221C subtilisin to select for improved peptide ligation. Ligase activity allowed product capture by the ligation of the subtilisin phagemids with improved ligase activity to a biotin-tagged peptide [3]. A second study successfully displayed fully active subtilisin on phage, although this involved addition of the subtilisin inhibitor CI2 to the culture medium. Selection for a change in P4 specificity then was carried out using a biotin-linked peptide diphenylester inhibitor [84].

3.3. Uncoupling prodomain processing from selection

A major limitation to any screening/selection method is that mutations affecting catalytic activity potentially affect the biosynthesis of subtilisin which is linked to autoprocessing of the prodomain [51]. Hence the selection of mutants will be biased toward enzymes which efficiently autoprocess. If the desired phenotype is activity toward a particular amino acid sequence, then the autoprocessing mechanism actually might be used to aid in selection by mutating the processing site on the prodomain to the target sequence [5,84]. This is apparently what occurred in the natural evolution of prohormone converting enzymes since the C-terminal sequence of the prodomain reflects the processing specificity [143]. If the desired phenotype is activity against a novel substrate, however, one needs to uncouple the biosynthesis of subtilisin from the selection for the new activity. This has been accomplished by using the 475–83 version of subtilisin, which is capable of folding without the prodomain [2,3,19,37].

3.4. Full circle

The first genetically engineered subtilisin appeared in the literature in 1985 and addressed the sensitivity of subtilisin to oxidation by peroxide [47]. It had been determined earlier that M222 is sensitive to oxidation leading to inactivation of the enzyme [146]. While it was clear that substituting for M222 would prevent this mechanism of inactivation by peroxide, it was not clear what amino acid would best substitute for methionine in providing optimal substrate interactions and preserving activity. For this reason, all 19 substitutions were made and the catalytic and stability properties of each compared. Thus even the first example of genetic-based protein engineering in subtilisin was in fact a random mutagenesis experiment which could be targeted to just one position because of detailed biochemical and structural information. After 15 years the best approach to 'engineering' desired properties into subtilisin probably remains targeted random mutagenesis, in which target selection is informed by all available information.

Acknowledgements

The author wishes to thank Patrick Alexander, Biao Ruan and Susan Strausberg for critically reading the manuscript. This study was supported by NIH grant GM42560.

References

- [1] L. Abrahmsen, J. Tom, J. Burnier, K.A. Butcher, A. Kosiakoff, J.A. Wells. Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution. *Biochemistry* 30 (1991) 4151–4159.
- [2] O. Almog, T. Gallagher, M. Tordova, J. Hoskins, P. Bryan, G.L. Gilliland. Crystal structure of calcium-independent subtilisin BPN' with restored thermal stability folded without the prodomain. *Proteins* 31 (1998) 21–32.
- [3] S. Atwell, J.A. Wells. Selection for improved subtiligases by phage display. *Proc. Natl. Acad. Sci. USA* 96 (1999) 9497–9502.
- [4] K.H. Bae, J.S. Jang, K.S. Park, S.H. Lee, S.M. Byun. Improvement of thermal stability of subtilisin J by changing the primary autolysis site. *Biochem. Biophys. Res. Commun.* 207 (1995) 20–24.
- [5] M.D. Ballinger, J. Tom, J.A. Wells. Designing subtilisin

- BPN' to cleave substrates containing dibasic residues, *Biochemistry* 34 (1995) 13312–13319.
- [6] M.D. Ballinger, J. Tom, J.A. Wells, Furilisin: a variant of subtilisin BPN' engineered for cleaving tribasic substrates, *Biochemistry* 35 (1996) 13579–13585.
 - [7] L.M. Bech, S. Branner, S. Hastrup, K. Breddam, Introduction of a free cysteinyl residue at position 68 in the subtilisin savinase, based on homology with proteinase K, *FEBS Lett.* 297 (1992) 164–166.
 - [8] L.M. Bech, S.B. Sorensen, K. Breddam, Mutational replacements in subtilisin 309. Val104 has a modulating effect on the P4 substrate preference, *Eur. J. Biochem.* 209 (1992) 869–874.
 - [9] L.M. Bech, S.B. Sorensen, K. Breddam, Significance of hydrophobic S4-P4 interactions in subtilisin 309 from *Bacillus lentus*, *Biochemistry* 32 (1993) 2845–2852.
 - [10] I.M. Bell, M.L. Fisher, Z.P. Wu, D. Hilvert, Kinetic studies on the peroxidase activity of selenosubtilisin, *Biochemistry* 32 (1993) 3754–3762.
 - [11] I.M. Bell, D. Hilvert, Peroxide dependence of the semisynthetic enzyme selenosubtilisin, *Biochemistry* 32 (1993) 13969–13973.
 - [12] A. Berger, I. Schechter, Mapping the active site of papain with the aid of peptide substrates and inhibitors, *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* 257 (1970) 249–264.
 - [13] C. Betzel, S. Kupsch, G. Papendorf, S. Hastrup, S. Branner, K.S. Wilson, Crystal structure of the alkaline proteinase savinase from *Bacillus lentus* at 1.4 Å resolution, *J. Mol. Biol.* 223 (1992) 427–445.
 - [14] W. Bode, E. Papamokos, D. Musil, The high-resolution x-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin C, an elastase inhibitor from the leech *Hirudo medicinalis*, *Eur. J. Biochem.* 166 (1987) 673–692.
 - [15] S. Braxton, J.A. Wells, The importance of a distal hydrogen bonding group in stabilizing the transition state in subtilisin BPN', *J. Biol. Chem.* 266 (1991) 11797–11800.
 - [16] S.B. Braxton, J.A. Wells, Incorporation of a stabilizing Ca-binding loop into subtilisin BPN', *Biochemistry* 31 (1992) 7796–7801.
 - [17] P.F. Brode 3rd, C.R. Erwin, D.S. Rauch, B.L. Barnett, J.M. Armprister, E.S. Wang, D.N. Rubingh, Subtilisin BPN' variants: increased hydrolytic activity on surface-bound substrates via decreased surface activity, *Biochemistry* 35 (1996) 3162–3169.
 - [18] P.F. Brode 3rd, C.R. Erwin, D.S. Rauch, D.S. Lucas, D.N. Rubingh, Enzyme behavior at surfaces. Site-specific variants of subtilisin BPN' with enhanced surface stability, *J. Biol. Chem.* 269 (1994) 23538–23543.
 - [19] P. Bryan, P. Alexander, S. Strausberg, F. Schwarz, L. Wang, G. Gilliland, D.T. Gallagher, Energetics of folding subtilisin BPN', *Biochemistry* 31 (1992) 4937–4945.
 - [20] P. Bryan, M.W. Pantoliano, S.G. Quill, H.Y. Hsiao, T. Poulos, Site-directed mutagenesis and the role of the oxyanion hole in subtilisin, *Proc. Natl. Acad. Sci. USA* 83 (1986) 3743–3745.
 - [21] P. Bryan, L. Wang, J. Hoskins, S. Ruvinov, S. Strausberg, P. Alexander, O. Almog, G. Gilliland, T.D. Gallagher, Catalysis of a protein folding reaction: mechanistic implications of the 2.0 Å structure of the subtilisin-prodomain complex, *Biochemistry* 34 (1995) 10310–10318.
 - [22] P.N. Bryan, in: T.J. Ahern, M.C. Manning (Eds.), *Pharmaceutical Biotechnology*, part B, Plenum Press, New York, 1992, pp. 147–181.
 - [23] P.N. Bryan, in: B.A. Shirley (Ed.), *Protein Stability and Folding: Theory and Practice*, vol. 40, Humana Press, Totowa, NJ, 1995, pp. 271–289.
 - [24] P.N. Bryan, in: U. Shinde, M. Inouye (Eds.), *Intramolecular Chaperones and Protein Folding*, R.G. Landes, Austin, TX, 1995, pp. 85–112.
 - [25] P.N. Bryan, M.P. Pantoliano, Combining Mutations for the Stabilization of Subtilisin, United States: Genex Corp., 1988.
 - [26] P.N. Bryan, M.L. Rollence, M.W. Pantoliano, J. Wood, B.C. Finzel, G.L. Gilliland, A.J. Howard, T.L. Poulos, Proteases of enhanced stability: characterization of a thermostable variant of subtilisin, *Proteins Struct. Funct. Genet.* 1 (1986) 326–334.
 - [27] P.N. Bryan, M.L. Rollence, J. Wood, S. Quill, S. Dodd, M. Whitlow, K. Hardman, M.W. Pantoliano, in: J. Gavora, D.F. Gerson, J. Luong, A. Storer, J.H. Woodley (Eds.), *Biotechnology Research and Applications*, Elsevier Applied Science Publ., Essex, 1988, pp. 57–67.
 - [28] P. Carter, L. Abrahmsen, J.A. Wells, Probing the mechanism and improving the rate of substrate-assisted catalysis in subtilisin BPN', *Biochemistry* 30 (1991) 6141–6148.
 - [29] P. Carter, B. Nilsson, J.P. Burnier, D. Burdick, J.A. Wells, Engineering subtilisin BPN' for site-specific proteolysis, *Proteins Struct. Funct. Genet.* 6 (1989) 240–248.
 - [30] P. Carter, J.A. Wells, Engineering enzyme specificity by 'substrate-assisted catalysis', *Science* 237 (1987) 394–399.
 - [31] P. Carter, J.A. Wells, Dissecting the catalytic triad of a serine protease, *Nature* 332 (1988) 564–568.
 - [32] P. Carter, J.A. Wells, Functional interaction among catalytic residues in subtilisin BPN', *Proteins Struct. Funct. Genet.* 7 (1990) 335–342.
 - [33] K. Chen, F.H. Arnold, Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide, *Proc. Natl. Acad. Sci. USA* 90 (1993) 5618–5622.
 - [34] N.M. Chu, Y. Chao, R.C. Bi, The 2 Å crystal structure of subtilisin E with PMSF inhibitor, *Protein Eng.* 8 (1995) 211–215.
 - [35] B.C. Cunningham, J.A. Wells, Improvement in the alkaline stability of subtilisin using an efficient random mutagenesis and screening procedure, *Protein Eng.* 1 (1987) 319–325.
 - [36] B.G. Davis, X. Shang, G. DeSantis, R.R. Bott, J.B. Jones, The controlled introduction of multiple negative charge at single amino acid sites in subtilisin *Bacillus lentus*, *Bioorg. Med. Chem.* 7 (1999) 2293–2301.
 - [37] S. Demartis, A. Huber, F. Viti, L. Lozzi, L. Giovannoni, P. Neri, G. Winter, D. Neri, A strategy for the isolation of

catalytic
phage.
[38] G. De
Jones.
modific
SI and
37 (199
[39] G. DeS
catalyti
at posit
Chem.
[40] G. DeS
specific
modific
specific
[41] D. Din
M. Mc
strate s
nol-sub
[42] J. Eder
BPN':
istry 32
[43] J. Eder
BPN':
293–304
[44] M.R. E
estein. J
charges
exchang
[45] C.R. Er
of engin
Protein
[46] D.A. Es
Burnier.
bic effec
neering.
[47] D.A. Es
zyme by
oxidatio
[48] C.O. Fa
Biochim.
[49] T.D. Ga
lin by d
213.
[50] T.D. Ga
zel (Eds.
tants E
Press, N.
[51] T.D. Ga
segment
cific fold
[52] N. Geno
Stability
Int. J. P
[53] D.W. G
Mielenz.

- catalytic activities from repertoires of enzymes displayed on phage, *J. Mol. Biol.* 286 (1999) 617–633.
- [38] G. DeSantis, P. Berglund, M.R. Stabile, M. Gold, J.B. Jones, Site-directed mutagenesis combined with chemical modification as a strategy for altering the specificity of the S1 and S1' pockets of subtilisin *Bacillus lentus*, *Biochemistry* 37 (1998) 5968–5973.
- [39] G. DeSantis, J.B. Jones, Probing the altered specificity and catalytic properties of mutant subtilisin chemically modified at position S156C and S166C in the S1 pocket, *Bioorg. Med. Chem.* 7 (1999) 1381–1387.
- [40] G. DeSantis, X. Shang, J.B. Jones, Toward tailoring the specificity of the S1 pocket of subtilisin *B. lentus*: chemical modification of mutant enzymes as a strategy for removing specificity limitations, *Biochemistry* 38 (1999) 13391–13397.
- [41] D. Dinakarpanthian, B.C. Shenoy, D. Hilvert, D.E. McRee, M. McTigue, P.R. Carey, Electric fields in active sites: substrate switching from null to strong fields in thiol- and seleno-subtilisins, *Biochemistry* 38 (1999) 6659–6667.
- [42] J. Eder, M. Rheinneckner, A.R. Fersht, Folding of subtilisin BPN': characterization of a folding intermediate, *Biochemistry* 32 (1993) 18–26.
- [43] J. Eder, M. Rheinneckner, A.R. Fersht, Folding of subtilisin BPN': role of the pro-sequence, *J. Mol. Biol.* 233 (1993) 293–304.
- [44] M.R. Egmond, W.P. Antheunisse, C.J. van Bommel, P. Ravestein, J. de Vlieg, H. Peters, S. Branner, Engineering surface charges in a subtilisin: the effects on electrophoretic and ion-exchange behaviour, *Protein Eng.* 7 (1994) 793–800.
- [45] C.R. Erwin, B.L. Barnett, J.D. Oliver, J.F. Sullivan, Effects of engineered salt bridges on the stability of subtilisin BPN', *Protein Eng.* 4 (1990) 87–97.
- [46] D.A. Estell, T.P. Graycar, J.V. Miller, D.B. Powers, J.P. Burnier, P.G. Ng, J.A. Wells, Probing steric and hydrophobic effects on enzyme-substrate interactions by protein engineering, *Science* 233 (1986) 659–663.
- [47] D.A. Estell, T.P. Graycar, J.A. Wells, Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation, *J. Biol. Chem.* 260 (1985) 6518–6521.
- [48] C.O. Fagain, Understanding and increasing protein stability, *Biochim. Biophys. Acta* 1252 (1995) 1–14.
- [49] T.D. Gallagher, P. Bryan, G. Gilliland, Calcium-free subtilisin by design, *Proteins Struct. Funct. Genet.* 16 (1993) 205–213.
- [50] T.D. Gallagher, G. Gilliland, P. Bryan, in: R. Bott, C. Betzel (Eds.), *Crystal Structure Analysis of Subtilisin BPN' Mutants Engineered for Studying Thermal Stability*, Plenum Press, New York, 1996.
- [51] T.D. Gallagher, G. Gilliland, L. Wang, P. Bryan, The prosegment-subtilisin BPN' complex: crystal structure of a specific foldase, *Structure* 3 (1995) 907–914.
- [52] N. Genov, B. Filippi, P. Dolashka, K.S. Wilson, C. Betzel, Stability of subtilisins and related proteinases (subtilases), *Int. J. Pept. Protein Res.* 45 (1995) 391–400.
- [53] D.W. Goddette, T. Christianson, B.F. Ladin, M. Lau, J.R. Mielenz, C. Paech, R.B. Reynolds, S.S. Yang, C.R. Wilson, Strategy and implementation of a system for protein engineering, *J. Biotechnol.* 28 (1993) 41–54.
- [54] T. Graycar, M. Knapp, G. Ganshaw, J. Dauberman, R. Bott, Engineered *Bacillus lentus* subtilisins having altered flexibility, *J. Mol. Biol.* 292 (1999) 97–109.
- [55] H. Gron, L.M. Bech, S. Branner, K. Breddam, A highly active and oxidation-resistant subtilisin-like enzyme produced by a combination of site-directed mutagenesis and chemical modification, *Eur. J. Biochem.* 194 (1990) 897–901.
- [56] H. Gron, L.M. Bech, S.B. Sorensen, M. Meldal, K. Breddam, Studies of binding sites in the subtilisin from *Bacillus lentus* by means of site directed mutagenesis and kinetic investigations, *Adv. Exp. Med. Biol.* 379 (1996) 105–112.
- [57] H. Gron, K. Breddam, Interdependency of the binding subsites in subtilisin, *Biochemistry* 31 (1992) 8967–8971.
- [58] H. Gron, M. Meldal, K. Breddam, Extensive comparison of the substrate preferences of two subtilisins as determined with peptide substrates which are based on the principle of intramolecular quenching, *Biochemistry* 31 (1992) 6011–6018.
- [59] P. Gros, K.H. Kalk, W.G.J. Hol, Calcium binding to thermolysin, *J. Biol. Chem.* 266 (1991) 2953–2961.
- [60] D. Haring, B. Hubert, E. Schuler, P. Schreier, Reasoning enantioselectivity and kinetics of seleno-subtilisin from the subtilisin template, *Arch. Biochem. Biophys.* 354 (1998) 263–269.
- [61] D. Haring, P. Schreier, From detergent additive to semisynthetic peroxidase-simplified and up-scaled synthesis of seleno-subtilisin, *Biotechnol. Bioeng.* 59 (1998) 786–791.
- [62] D. Haring, P. Schreier, Chemical engineering of enzymes: altered catalytic activity, predictable selectivity and exceptional stability of the semisynthetic peroxidase seleno-subtilisin, *Naturwissenschaften* 86 (1999) 307–312.
- [63] D. Haring, E. Schuler, A. Waldemar, C.R. Saha-Moller, P. Schreier, Semisynthetic enzymes in asymmetric synthesis: enantioselective reduction of racemic hydroperoxides catalyzed by seleno-subtilisin, *J. Org. Chem.* 64 (1999) 832–835.
- [64] T. Hayashi, M. Matsubara, D. Nohara, S. Kojima, K. Miura, T. Sakai, Renaturation of the mature subtilisin BPN' immobilized on agarose beads, *FEBS Lett.* 350 (1994) 109–112.
- [65] J. Heringa, P. Argos, M.R. Egmond, J. de Vlieg, Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters, *Protein Eng.* 8 (1995) 21–30.
- [66] H. Hirohara, M. Philipp, M.L. Bender, Binding rates, O-S substitution effects, and the pH dependence of chymotrypsin reactions, *Biochemistry* 16 (1977) 1573–1580.
- [67] Z. Hu, K. Haghighi, F. Jordan, Further evidence for the structure of the subtilisin propeptide and for its interactions with mature subtilisin, *J. Biol. Chem.* 271 (1996) 3375–3384.
- [68] Z. Hu, X. Zhu, F. Jordan, M. Inouye, A covalently trapped folding intermediate of subtilisin E: spontaneous dimerization of a prosubtilisin E Ser49Cys mutant in vivo and its autoprocesing in vitro, *Biochemistry* 33 (1994) 562–569.
- [69] W. Huang, J. Wang, D. Bhattacharyya, L.G. Bachas,

- Improving the activity of immobilized subtilisin by site-specific attachment to surfaces, *Anal. Chem.* 69 (1997) 4601–4607.
- [70] A. Ikai, Denaturation of subtilisin BPN' and its derivatives in aqueous guanidine hydrochloride solutions, *Biochim. Biophys. Acta* 445 (1976) 182–193.
- [71] H. Ikemura, H. Takagi, M. Inouye, Requirement of pro sequence for the production of active subtilisin in *Escherichia coli*, *J. Biol. Chem.* 262 (1987) 7859–7864.
- [72] M. Jacobs, M. Eliason, M. Uhlen, J. Flock, Cloning, sequencing and expression of subtilisin Carlsberg from *Bacillus licheniformis*, *Nucleic Acids Res.* 13 (1985) 8913–8926.
- [73] S.C. Jain, U. Shinde, Y. Li, M. Inouye, H.M. Berman, The crystal structure of an autoprocessed Ser221Cys-subtilisin E-propeptide complex at 2.0 Å resolution, *J. Mol. Biol.* 284 (1998) 137–144.
- [74] J.S. Jang, K.H. Bae, S.M. Byun, Effect of the weak Ca(2+)-binding site of subtilisin J by site-directed mutagenesis on heat stability, *Biochem. Biophys. Res. Commun.* 188 (1992) 184–189.
- [75] J.S. Jang, D.K. Park, M. Chun, S.M. Byun, Identification of autoproteolytic cleavage site in the Asp-49 mutant subtilisin J by site-directed mutagenesis, *Biochim. Biophys. Acta* 1162 (1993) 233–235.
- [76] L.J. Jensen, K.V. Andersen, A. Svendsen, T. Kretschmar, Scoring functions for computational algorithms applicable to the design of spiked oligonucleotides, *Nucleic Acids Res.* 26 (1998) 697–702.
- [77] H. Kano, S. Taguchi, H. Momose, Cold adaptation of a mesophilic serine protease, subtilisin, by in vitro random mutagenesis, *Appl. Microbiol. Biotechnol.* 47 (1997) 46–51.
- [78] T.W. Keough, Y. Sun, B.L. Barnett, M.P. Lacey, M.D. Bauer, E.S. Wang, C.R. Erwin, Rapid analysis of single-cysteine variants of recombinant proteins, *Methods Mol. Biol.* 61 (1996) 171–183.
- [79] R.D. Kidd, P. Sears, D.H. Huang, K. Witte, C.H. Wong, G.K. Farber, Breaking the low barrier hydrogen bond in a serine protease, *Protein Sci.* 8 (1999) 410–417.
- [80] R.D. Kidd, H.P. Yennawar, P. Sears, C.-H. Wong, G.K. Farber, A weak calcium binding site in subtilisin BPN' has a dramatic effect on protein stability, *J. Am. Chem. Soc.* 118 (1996) 1645–1650.
- [81] A. Knappik, L. Ge, A. Honegger, P. Pack, M. Fischer, G. Wellenhofer, A. Hoess, J. Wolle, A. Pluckthun, B. Virnekas, Fully synthetic human combinatorial antibody libraries (Hu-CAL) based on modular consensus frameworks and CDRs randomized with trinucleotides, *J. Mol. Biol.* 296 (2000) 57–86.
- [82] T. Kobayashi, M. Inouye, Functional analysis of the intramolecular chaperone. Mutational hot spots in the subtilisin pro-peptide and a second site suppressor mutation within the subtilisin molecule, *J. Mol. Biol.* 226 (1992) 931–933.
- [83] W. Kullman, *Enzymatic Peptide Synthesis*, CRC Press, Boca Raton, FL, 1987.
- [84] D. Legendre, N. Laraki, T. Graslund, M.E. Bjornvad, M. Bouchet, P.A. Nygren, T.V. Borchert, J. Fastrez, Display of active subtilisin 309 on phage: analysis of parameters influencing the selection of subtilisin variants with changed substrate specificity from libraries using phosphonylating inhibitors, *J. Mol. Biol.* 296 (2000) 87–102.
- [85] J.P. Leis, C.E. Cameron, Engineering proteases with altered specificity, *Curr. Opin. Biotechnol.* 5 (1994) 403–408.
- [86] Y. Li, Z. Hu, F. Jordan, M. Inouye, Functional analysis of the propeptide of subtilisin E as an intramolecular chaperone for protein folding. Refolding and inhibitory abilities of propeptide mutants, *J. Biol. Chem.* 270 (1995) 25127–25132.
- [87] Y. Li, M. Inouye, Autoprocessing of prothiolsubtilisin E in which active-site serine 221 is altered to cysteine, *J. Biol. Chem.* 269 (1994) 4169–4174.
- [88] Y. Li, M. Inouye, The mechanism of autoprocessing of the propeptide of prosubtilisin E: intramolecular or intermolecular event?, *J. Mol. Biol.* 262 (1996) 591–594.
- [89] W. Lu, I. Apostol, M.A. Qasim, N. Warne, R. Wynn, W.L. Zhang, S. Anderson, Y.W. Chiang, E. Ogin, I. Rothberg, K. Ryan, M. Laskowski Jr., Binding of amino acid side-chains to S1 cavities of serine proteinases, *J. Mol. Biol.* 266 (1997) 441–461.
- [90] K. Masuda-Momma, T. Hatanaka, K. Inouye, K. Kanaori, A. Tamura, K. Akasaka, S. Kojima, I. Kumagai, K. Miura, B. Tonomura, Interaction of subtilisin BPN' and recombinant *Streptomyces* subtilisin inhibitors with substituted P1 site residues, *J. Biochem.* 114 (1993) 553–559.
- [91] K. Masuda-Momma, T. Shimakawa, K. Inouye, K. Hiromi, S. Kojima, I. Kumagai, K. Miura, B. Tonomura, Identification of amino acid residues responsible for the changes of absorption and fluorescence spectra on the binding of subtilisin BPN' and *Streptomyces* subtilisin inhibitor, *J. Biochem.* 114 (1993) 906–911.
- [92] M. Matsubara, E. Kurimoto, S. Kojima, K. Miura, T. Sakai, Achievement of renaturation of subtilisin BPN' by a novel procedure using organic salts and a digestible mutant of *Streptomyces* subtilisin inhibitor, *FEBS Lett.* 342 (1994) 193–196.
- [93] C.A. McPhalen, M.N.G. James, Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-C-subtilisin Carlsberg and Cl-2-subtilisin novo, *Biochemistry* 27 (1988) 6582–6598.
- [94] H.C. Mei, Y.C. Liaw, Y.C. Li, D.C. Wang, H. Takagi, Y.C. Tsai, Engineering subtilisin Yab: restriction of substrate specificity by the substitution of Gly124 and Gly151 with Ala, *Protein Eng.* 11 (1998) 109–117.
- [95] C. Mitchinson, J.A. Wells, Protein engineering of disulfide bonds in subtilisin BPN', *Biochemistry* 28 (1989) 4807–4815.
- [96] K. Miyazaki, F.H. Arnold, Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function, *J. Mol. Evol.* 49 (1999) 716–720.
- [97] N. Mizushima, D. Spellmeyer, S. Hirono, D. Pearlman, P. Kollman, Free energy perturbation calculations on binding and catalysis after mutating threonine 220 in subtilisin, *J. Biol. Chem.* 266 (1991) 11801–11809.
- [98] T. Nakatsuka, T. Sasaki, E.T. Kaiser, Peptide segment cou
- pling
J. A
[99] L.O.
S. Fi
ski.
tions
[100] E. N
phili
muta
to co
[101] K.E.
the a
tion.
[102] K.E.
ol-sul
ine re
243 (1
[103] J.E.
T.V.
of su
(1999
[104] T.P. (1
chin,
Asn-1
dral a
[105] D. O
Struct
9B (1
[106] C.N.
Confe
with :
Chem.
[107] C. Pa
Unusu
engine
379 (1
[108] M.P. :
lized I
[109] M.W.
ronme
Struct.
[110] M.W.
ence, J
tilisin I
cystein
2077–2
[111] M.W.
K.D.
creases
cremen
istry 2:
[112] M.W. :
B.C. I
engine:
for the
Bioche:

- plung catalyzed by the semisynthetic enzyme thiolsubtilisin, *J. Am. Chem. Soc.* 109 (1987) 3808–3810.
- [99] L.O. Narhi, Y. Stabinsky, M. Levitt, L. Miller, R. Sachdev, S. Finley, S. Park, C. Kolvenbach, T. Arakawa, M. Zukowski, Enhanced stability of subtilisin by three point mutations, *Biotechnol. Appl. Biochem.* 13 (1991) 12–24.
- [100] E. Narinx, E. Baise, C. Gerday, Subtilisin from psychrophilic antarctic bacteria: characterization and site-directed mutagenesis of residues possibly involved in the adaptation to cold, *Protein Eng.* 10 (1997) 1271–1279.
- [101] K.E. Neet, D.E. Koshland Jr., The conversion of serine at the active site of subtilisin to cysteine: a 'chemical mutation', *Proc. Natl. Acad. Sci. USA* 56 (1966) 1606–1611.
- [102] K.E. Neet, A. Nanci, D.E. Koshland Jr., Properties of thiol-subtilisin. The consequences of converting the active serine residue to cysteine in a serine protease, *J. Biol. Chem.* 243 (1968) 6392–6401.
- [103] J.E. Ness, M. Welch, L. Giver, M. Bueno, J.R. Cherry, T.V. Borchert, W.P. Stemmer, J. Minshull, DNA shuffling of subgenomic sequences of subtilisin, *Nat. Biotechnol.* 17 (1999) 893–896.
- [104] T.P. O'Connell, R.M. Day, E.V. Torchilin, W.W. Bachovchin, J.G. Malthouse, A ¹³C-NMR study of the role of Asn-155 in stabilizing the oxyanion of a subtilisin tetrahedral adduct, *Biochem. J.* 326 (1997) 861–866.
- [105] D. Oxender (Organizer), UCLA Symposium on Protein Structure, Folding and Design, *J. Cell. Biochem. Suppl.* 9B (1985) 91–145.
- [106] C.N. Pace, G.R. Grimsley, J.A. Thomson, B.J. Barnett, Conformational stabilities and activity of ribonuclease T1 with zero, one and two intact disulfide bonds, *J. Biol. Chem.* 263 (1988) 11820–11825.
- [107] C. Paech, D.W. Goddette, T. Christianson, C.R. Wilson, Unusual ligand binding at the active site domain of an engineered mutant of subtilisin BL, *Adv. Exp. Med. Biol.* 379 (1996) 257–268.
- [108] M.P. Pantoliano, R.C. Ladner, Computer Designed Stabilized Proteins, United States: Genex Corp., 1987.
- [109] M.W. Pantoliano, Proteins designed for challenging environments and catalysis in organic solvents, *Curr. Opin. Struct. Biol.* 2 (1992) 559–568.
- [110] M.W. Pantoliano, R.C. Ladner, P.N. Bryan, M.L. Rollence, J.F. Wood, T.L. Poulos, Protein engineering of subtilisin BPN': stabilization through the introduction of two cysteines to form a disulfide bond, *Biochemistry* 26 (1987) 2077–2082.
- [111] M.W. Pantoliano, M. Whitlow, J.F. Wood, S.W. Dodd, K.D. Hardman, M.L. Rollence, P.N. Bryan, Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding, *Biochemistry* 28 (1989) 7205–7213.
- [112] M.W. Pantoliano, M. Whitlow, J.F. Wood, M.L. Rollence, B.C. Finzel, G. Gilliland, T.L. Poulos, P.N. Bryan, The engineering of binding affinity at metal ion binding sites for the stabilization of proteins: subtilisin as a test case, *Biochemistry* 27 (1988) 8311–8317.
- [113] J.J. Perona, C.S. Craik, Structural basis of substrate specificity in the serine proteases, *Protein Sci.* 4 (1995) 337–360.
- [114] E.B. Peterson, D. Hilvert, Nonessential active site residues modulate selenosubtilisin's kinetic mechanism, *Biochemistry* 34 (1995) 6616–6620.
- [115] M. Philipp, M.L. Bender, Kinetics of subtilisin and thiol-subtilisin, *Mol. Cell. Biochem.* 51 (1983) 5–32.
- [116] M. Philipp, I.H. Tsai, M.L. Bender, Comparison of the kinetic specificity of subtilisin and thiolsubtilisin toward *n*-alkyl *p*-nitrophenyl esters, *Biochemistry* 18 (1979) 3769–3773.
- [117] E. Plettner, G. DeSantis, M.R. Stabile, J.B. Jones, Modulation of esterase and amidase activity of subtilisin *Bacillus lentus* by chemical modification of cysteine mutants, *J. Am. Chem. Soc.* 121 (1999) 4977–4981.
- [118] D.C. Poland, H.A. Scheraga, Statistical mechanics of non-covalent bonds in polyamino acids. VIII. Covalent loops in proteins, *Biopolymers* 3 (1965) 379–399.
- [119] L. Polgar, M.L. Bender, The reactivity of thiol-subtilisin, an enzyme containing a synthetic functional group, *Biochemistry* 6 (1967) 610–620.
- [120] L. Polgar, M.L. Bender, Chromatography and activity of thiol-subtilisin, *Biochemistry* 8 (1969) 136–141.
- [121] S.N. Rao, U.C. Singh, P.A. Bash, P.A. Kollman, Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin, *Nature* 328 (1987) 551–554.
- [122] M. Rheinhecker, G. Baker, J. Eder, A.R. Fersht, Engineering a novel specificity in subtilisin BPN', *Biochemistry* 32 (1993) 1199–1203.
- [123] M. Rheinhecker, J. Eder, P.S. Pandey, A.R. Fersht, Variants of subtilisin BPN' with altered specificity profiles, *Biochemistry* 33 (1994) 221–225.
- [124] M.L. Rollence, D. Filpula, M.W. Pantoliano, P.N. Bryan, Engineering thermostability in subtilisin BPN' by in vitro mutagenesis, *CRC Crit. Rev. Biotechnol.* 8 (1988) 217–224.
- [125] I.S. Rombauer, The Joy of Cooking, 1931.
- [126] B. Ruan, Folding of Subtilisin: Study of Independent Folding and Pro-domain Catalyzed Folding, Ph.D. Dissertation, University of Maryland, College Park, MD, 1998.
- [127] B. Ruan, J. Hoskins, P.N. Bryan, Rapid folding of calcium-free subtilisin by a stabilized pro-domain mutant, *Biochemistry* 38 (1999) 8562–8571.
- [128] B. Ruan, J. Hoskins, L. Wang, P.N. Bryan, Stabilizing the subtilisin BPN' pro-domain by phage display selection: how restrictive is the amino acid code for maximum protein stability? [In process citation], *Protein Sci.* 7 (1998) 2345–2353.
- [129] A.J. Russell, A.R. Fersht, Rational modification of enzyme catalysis by engineering surface charge, *Nature* 328 (1987) 496–500.
- [130] A.J. Russell, P.G. Thomas, A.R. Fersht, Electrostatic effects on modification of charged groups in the active site cleft of subtilisin by protein engineering, *J. Mol. Biol.* 193 (1987) 803–813.
- [131] S. Ruvinov, L. Wang, B. Ruan, O. Almog, G. Gilliland, E.

- Eisenstein, P. Bryan, Engineering the independent folding of the subtilisin BPN' prodomain: analysis of two-state folding vs. protein stability. *Biochemistry* 36 (1997) 10414–10421.
- [132] A. Sattler, S. Kanka, K.H. Maurer, D. Riesner, Thermally stable variants of subtilisin selected by temperature-gradient gel electrophoresis, *Electrophoresis* 17 (1996) 784–792.
- [133] R. Schulein, J. Kreft, S. Gonski, W. Goebel, Preprosubtilisin Carlsberg processing and secretion is blocked after deletion of amino acids 97–101 in the mature part of the enzyme, *Mol. Gen. Genet.* 227 (1991) 137–143.
- [134] P. Sears, M. Schuster, P. Wang, K. Witte, C.-H. Wong, Engineering subtilisin for peptide coupling: studies on the effects of counterions and site-specific modifications on the stability and specificity of the enzyme, *J. Am. Chem. Soc.* 116 (1994) 6521–6530.
- [135] S. Shafikhani, R.A. Siegel, E. Ferrari, V. Schellenberger, Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization, *BioTechniques* 23 (1997) 304–310.
- [136] Z. Shao, F.H. Arnold, Engineering new functions and altering existing functions, *Curr. Opin. Struct. Biol.* 6 (1996) 513–518.
- [137] Z. Shao, H. Zhao, L. Giver, F.H. Arnold, Random-priming in vitro recombination: an effective tool for directed evolution, *Nucleic Acids Res.* 26 (1998) 681–683.
- [138] U. Shinde, X. Fu, M. Inouye, A pathway for conformational diversity in proteins mediated by intramolecular chaperones, *J. Biol. Chem.* 274 (1999) 15615–15621.
- [139] U. Shinde, M. Inouye, Folding mediated by an intramolecular chaperone: autoprocessing pathway of the precursor resolved via a substrate assisted catalysis mechanism, *J. Mol. Biol.* 247 (1995) 390–395.
- [140] U. Shinde, M. Inouye, Folding pathway mediated by an intramolecular chaperone: characterization of the structural changes in pro-subtilisin E coincident with autoprocessing, *J. Mol. Biol.* 252 (1995) 25–30.
- [141] U. Shinde, M. Inouye, Propeptide-mediated folding in subtilisin: the intramolecular chaperone concept, *Adv. Exp. Med. Biol.* 379 (1996) 147–154.
- [142] U.P. Shinde, J.J. Liu, M. Inouye, Protein memory through altered folding mediated by intramolecular chaperones, *Nature* 389 (1997) 520–522.
- [143] R.J. Siezen, J.A.M. Leunissen, U. Shinde, in: U. Shinde, M. Inouye (Eds.), *Intramolecular Chaperones and Protein Folding*, R.G. Landes, Austin, TX, 1995, pp. 233–256.
- [144] S.B. Sorensen, L.M. Bech, M. Meldal, K. Breddam, Mutational replacements of the amino acid residues forming the hydrophobic S4 binding pocket of subtilisin 309 from *Bacillus lentus*, *Biochemistry* 32 (1993) 8994–8999.
- [145] R. Sowdhamini, N. Srinivasan, B. Shoichet, D.V. Santi, C. Ramakrishnan, P. Balaram, Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis, *Protein Eng.* 3 (1989) 95–103.
- [146] C.E. Stauffer, D. Etson, The effect on subtilisin activity of oxidizing a methionine residue, *J. Biol. Chem.* 244 (1969) 5333–5338.
- [147] M.J. Sternberg, F.R. Hayes, A.J. Russell, P.G. Thomas, A.R. Fersht, Prediction of electrostatic effects of engineering of protein charges, *Nature* 330 (1987) 86–88.
- [148] M.J.E. Sternberg, F.R.F. Hayes, A.J. Russell, P.G. Thomas, A.R. Fersht, Prediction of electrostatic effects of engineering of protein charges, *Nature* 330 (1987) 86–88.
- [149] S. Strausberg, P. Alexander, D.T. Gallagher, G. Gilliland, B.L. Barnett, P. Bryan, Directed evolution of a subtilisin with calcium-independent stability, *Bio/technology* 13 (1995) 669–673.
- [150] S. Strausberg, P. Alexander, L. Wang, D.T. Gallagher, G. Gilliland, P. Bryan, An engineered disulfide crosslink accelerates the refolding rate of calcium-free subtilisin by 850-fold, *Biochemistry* 32 (1993) 10371–10377.
- [151] S. Strausberg, P. Alexander, L. Wang, F. Schwarz, P. Bryan, Catalysis of a protein folding reaction: thermodynamic and kinetic analysis of subtilisin BPN' interactions with its propeptide fragment, *Biochemistry* 32 (1993) 8112–8119.
- [152] R. Syed, Z.P. Wu, J.M. Hogle, D. Hilvert, Crystal structure of selenosubtilisin at 2.0-Å resolution, *Biochemistry* 32 (1993) 6157–6164.
- [153] S. Taguchi, A. Ozaki, H. Momose, Engineering of a cold-adapted protease by sequential random mutagenesis and a screening system, *Appl. Environ. Microbiol.* 64 (1998) 492–495.
- [154] S. Taguchi, A. Ozaki, T. Nonaka, Y. Mitsui, H. Momose, A cold-adapted protease engineered by experimental evolution system, *J. Biochem.* 126 (1999) 689–693.
- [155] H. Takagi, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [156] H. Takagi, T. Maeda, I. Ohtsu, Y.C. Tsai, S. Nakamori, Restriction of substrate specificity of subtilisin E by introduction of a side chain into a conserved glycine residue, *FEBS Lett.* 395 (1996) 127–132.
- [157] H. Takagi, Y. Morinaga, H. Ikemura, M. Inouye, Mutant subtilisin E with enhanced protease activity obtained by site-directed mutagenesis, *J. Biol. Chem.* 263 (1988) 19592–19596.
- [158] H. Takagi, Y. Morinaga, H. Ikemura, M. Inouye, The role of Pro-239 in the catalysis and heat stability of subtilisin E, *J. Biochem.* 105 (1989) 953–956.
- [159] H. Takagi, I. Ohtsu, S. Nakamori, Construction of novel subtilisin E with high specificity, activity and productivity through multiple amino acid substitutions, *Protein Eng.* 10 (1997) 985–989.
- [160] H. Takagi, T. Takahashi, H. Momose, M. Inouye, Y. Maeda, H. Matsuzawa, T. Ohta, Enhancement of the thermostability of subtilisin E by introduction of a disulfide bond engineered on the basis of structural comparison with a thermophilic subtilisin, *Biochemistry* 32 (1993) 6874–6877.
- [161] H. Takagi, M. Inouye, P. Bryan, Directed evolution of a subtilisin with calcium-independent stability, *Bio/technology* 13 (1995) 669–673.
- [162] K. Takagi, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [163] T. Tanaka, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [164] T. Tanaka, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [165] T. Tanaka, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [166] T. Tange, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [167] A.V. Teplov, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [168] P.G. Thorpe, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [169] P.J. Tong, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [170] K.M. Ulnowski, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [171] N. Vasanthakumari, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [172] A. Volkov, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [173] A.A. Volkov, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.
- [174] C. von der Mühlen, S. Arafuka, M. Inouye, M. Yamasaki, The effect of amino acid deletion in subtilisin E, based on structural comparison with a microbial alkaline elastase, on its substrate specificity and catalysis, *J. Biochem.* 111 (1992) 584–588.

- 44 (1969)
- Thomas, engineer.
- P.G. Thomas, effects of enzyme 36–88.
- Gilliland, subtilisin biology 13
- Gallagher, G. link acceleration by 850.
- Warz, P. thermodynamic interactions 93) 8112–
- structure ministry 32
- of a cold-sis and a 998) 492–
- Momose, tal evolution.
- i. The effect on structure, on its 11 (1992)
- Nakamori, by introduction of residue.
- i. Mutant obtained by 3 (1988)
- The role of subtilisin E.
- of novel ductivity i Eng. 10
- Y. Mae, thermodynamic bond with *
- thermophilic serine protease, *J. Biol. Chem.* 265 (1990) 6874–6878.
- [161] H. Takagi, M. Yamamoto, I. Ohtsu, S. Nakamori, Random mutagenesis into the conserved Gly154 of subtilisin E: isolation and characterization of the revertant enzymes, *Protein Eng.* 11 (1998) 1205–1210.
- [162] K. Takahashi, J.M. Sturtevant, Thermal denaturation of streptomyces subtilisin inhibitor, subtilisin BPN', and the inhibitor-subtilisin complex, *Biochemistry* 20 (1981) 6185–6190.
- [163] T. Tanaka, H. Matsuzawa, S. Kojima, I. Kumagai, K. Miura, T. Ohta, P1 specificity of aqualysin I (a subtilisin-type serine protease) from *Thermus aquaticus* YT-1, using P1-substituted derivatives of Streptomyces subtilisin inhibitor, *Biosci. Biotechnol. Biochem.* 62 (1998) 2035–2038.
- [164] T. Tanaka, H. Matsuzawa, T. Ohta, Engineering of S2 site of aqualysin I: alteration of P2 specificity by excluding P2 side chain, *Biochemistry* 37 (1998) 17402–17407.
- [165] T. Tanaka, H. Matsuzawa, T. Ohta, Identification and designing of the S3 site of aqualysin I, a thermophilic subtilisin-related serine protease, *J. Biochem.* 125 (1999) 1016–1021.
- [166] T. Tange, S. Taguchi, S. Kojima, K. Miura, H. Momose, Improvement of a useful enzyme (subtilisin BPN') by an experimental evolution system, *Appl. Microbiol. Biotechnol.* 41 (1994) 239–244.
- [167] A.V. Teplyakov, J.M. van der Laan, A.A. Lammers, H. Kelders, K.H. Kalk, O. Misser, L.J. Mulleners, B.W. Dijkstra, Protein engineering of the high-alkaline serine protease PB92 from *Bacillus alcalophilus*: functional and structural consequences of mutation at the S4 substrate binding pocket, *Protein Eng.* 5 (1992) 413–420.
- [168] P.G. Thomas, A.J. Russell, A.R. Fersht, Tailoring the pH dependence of enzyme catalysis using protein engineering, *Nature* 318 (1985) 375–376.
- [169] P.J. Tonge, P.R. Carey, Length of the acyl carbonyl bond in acyl-serine proteases correlates with reactivity, *Biochemistry* 29 (1990) 10723–10727.
- [170] K.M. Ulmer, Protein engineering, *Science* 219 (1983) 666–671.
- [171] N. Vasantha, L.D. Thompson, C. Rhodes, C. Banner, J. Nagle, D. Filpula, Genes for alkaline and neutral protease from *Bacillus amyloliquefaciens* contain a large open-reading frame between the regions coding for signal sequence and mature protein, *J. Bacteriol.* 159 (1984) 811–819.
- [172] A. Volkov, F. Jordan, Evidence for intramolecular processing of prosubtilisin sequestered on a solid support, *J. Mol. Biol.* 262 (1996) 595–599.
- [173] A.A. Volkov, Z. Shao, F.H. Arnold, Recombination and chimeraesis by in vitro heteroduplex formation and in vivo repair, *Nucleic Acids Res.* 27 (1999) e18.
- [174] C. von der Osten, S. Branner, S. Hastrup, L. Hedegaard, M.D. Rasmussen, H. Bisgaard-Frantzen, S. Carlsen, J.M. Mikkelsen, Protein engineering of subtilisins to improve stability in detergent formulations, *J. Biotechnol.* 28 (1993) 55–68.
- [175] G. Voordouw, C. Milo, R.S. Roche, Role of bound calcium in thermostable, proteolytic enzymes. Separation of intrinsic and calcium ion contributions to the kinetic thermal stability, *Biochemistry* 15 (1976) 3716–3724.
- [176] L. Wang, B. Ruan, S. Ruvinov, P.N. Bryan, Engineering the independent folding of the subtilisin BPN' pro-domain: correlation of pro-domain stability with the rate of subtilisin folding, *Biochemistry* 37 (1998) 3165–3171.
- [177] L. Wang, S. Ruvinov, S. Strausberg, T.D. Gallagher, G. Gilliland, P. Bryan, Prodomain mutations at the subtilisin interface: correlation of binding energy and the rate of catalyzed folding, *Biochemistry* 34 (1995) 15415–15420.
- [178] P.P. Wangikar, J.O. Rich, D.S. Clark, J.S. Dordick, Probing enzymic transition state hydrophobicities, *Biochemistry* 34 (1995) 12302–12310.
- [179] J.A. Wells, Additivity of mutational effects in proteins, *Biochemistry* 29 (1990) 8509–8517.
- [180] J.A. Wells, B.C. Cunningham, T.P. Graycar, D.A. Estell, Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin, *Philos. Trans. R. Soc. London* 317 (1986) 415–423.
- [181] J.A. Wells, B.C. Cunningham, T.P. Graycar, D.A. Estell, Recruitment of substrate-specificity properties from one enzyme into a related one by protein engineering, *Proc. Natl. Acad. Sci. USA* 84 (1987) 5167–5171.
- [182] J.A. Wells, E. Ferrari, D.J. Henner, D.A. Estell, E.Y. Chen, Cloning, sequencing and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillus subtilis*, *Nucleic Acids Res.* 11 (1983) 7911–7925.
- [183] J.A. Wells, D.B. Powers, In vivo formation and stability of engineered disulfide bonds in subtilisin, *J. Biol. Chem.* 261 (1986) 6564–6570.
- [184] J.A. Wells, D.B. Powers, R.R. Bott, T.P. Graycar, D.A. Estell, Designing substrate specificity by protein engineering of electrostatic interactions, *Proc. Natl. Acad. Sci. USA* 84 (1987) 1219–1223.
- [185] A.K. Whiting, W.L. Peticolas, Details of the acyl-enzyme intermediate and the oxyanion hole in serine protease catalysis, *Biochemistry* 33 (1994) 552–561.
- [186] C.-H. Wong, S.-T. Chen, W.J. Hennen, J.A. Bibbs, Y.-F. Wang, J.L.-C. Liu, M.W. Pantoliano, M. Whitlow, P.N. Bryan, Enzymes in organic synthesis: use of subtilisin and a highly stable mutant derived from multiple site-specific mutations, *J. Am. Chem. Soc.* 112 (1990) 945–953.
- [187] C.H. Wong, Enzymatic catalysts in organic synthesis, *Science* 244 (1989) 1145–1152.
- [188] C.H. Wong, G.J. Shen, R.L. Pederson, Y.F. Wang, W.J. Hennen, Enzymatic catalysis in organic synthesis, *Methods Enzymol.* 202 (1991) 591–620.
- [189] L. You, F.H. Arnold, Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide, *Protein Eng.* 9 (1996) 77–83.
- [190] H. Zhao, F.H. Arnold, Functional and nonfunctional mutations distinguished by random recombination of homologous genes, *Proc. Natl. Acad. Sci. USA* 94 (1997) 7997–8000.

- [191] H. Zhao, F.H. Arnold, Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein Eng.* 12 (1999) 47–53.
- [192] H. Zhao, L. Giver, Z. Shao, J.A. Affholter, F.H. Arnold, Molecular evolution by staggered extension process (StEP) in vitro recombination [see comments]. *Nat. Biotechnol.* 16 (1998) 258–261.
- [193] H. Zhao, Y. Li, F.H. Arnold, Strategy for the directed evolution of a peptide ligase. *Ann. NY Acad. Sci.* 799 (1996) 1–5.
- [194] L. Zhu, Y. Ji, Protein engineering on subtilisin E. *Chin. J. Biotechnol.* 13 (1997) 9–15.



Keywords

1. Intro

Lipase
or form
present
and is cl
is show

Triglycer

The lipa
often ex
pase, ly
nase, an
[1,2,27].
substrate
ceride, a
ceride as
acids alc
for the p
sn2 is de
zymes co
overall s
tility of
and exhi
definition
which as
thus incl

• Corre
E-mail: as

0167-4838
PII: S016

Exhibit 8

JOURNAL OF SCIENCE

THE FRANCIS A. COLE
LIBRARY OF MEDICINE
BOSTON, MA

APR 14 1988

NM19160904DEC88 80408602P38 14
FRANCIS COUNTWAY LIB MED DIR
10 SHATTUCK ST
COPY 1
BOSTON MA 02115

nature

7 April 1988

Vol. 332 Issue no. 6164

A view across sand dunes in the Sahara. A study of wind-driven sand transport in the north-western Algerian Sahara identifies a previously unrecognized mechanism, page 532. (Photo: Frank Lane.)

THIS WEEK

Crime-fighting advance
Using the DNA polymerase chain reaction, DNA can now be typed from a single hair. As hairs are one of the most frequent forms of evidence at scenes of crimes the consequences for forensic science are considerable, page 543.



Ring nebula cycle
Twenty years after they were predicted, a new class of cosmological X-ray source is discovered. Ring nebula NGC6888 is the first, pages 518 and 486.

Resistance evasion
A bacterial pathogen of the pepper plant that has mutated to evade host recognition has a transposable element in a gene responsible for the plant's hypersensitive response, page 541.

'Greenhouse' gas rising
Levels of atmospheric methane, a candidate for contributing to global warming, are increasing. Radiocarbon data suggest that over 30 per cent of atmospheric methane is derived from fossil carbon, pages 522 and 489.

Developmental switch
The switch from mitosis to meiosis in yeast has been pinned down to the inhibition of a protein kinase by a product of a gene specifically activated in diploid cells, page 509.

Lochs more bonnie

Have reductions in sulphur emissions and acid rain deposition in the past decades led to improvements in the environment? Chemical and diatom analyses of a pair of Scottish lochs give some of the answers, page 530.

Brain power

Electron microscopy shows the brain protein MAP 1C, thought to be responsible for the transport of cytoplasmic organelles, to be structurally similar to dynein, the force-generating protein in cilia and flagella. See page 561.

Titanic collisions

Earthly laboratory experiments provide evidence to support the idea that the nitrogen gas present on Saturn's moon Titan formed from ammonia as a result of high-velocity collisions with meteors, page 520.

Great Lakes battle

Despite an 'invasion' from the north by a voracious predator, the factors limiting the algal biomass in Lake Michigan seem to be related to nutrient supply, not a prey/predator balance, pages 537 and 491.

Guide to Authors

Facing page 568.

OPINION

APR 1 1988

A united Europe in 1992? ■ Squaring the circle
Windows copyright? 473

NEWS

US-Japan agreement ■ AIDS drug ■ Australian reform ■
Space settlement ■ Congress and NIH ■ Armenia/
Azerbaijan ■ Sequencing yeast genome ■ UK wind power
■ Simultaneous classroom ■ UK defence spending ■
Superconductors ■ Sea to pond in Japan ■ Leningrad
library fire ■ Correspondence 475-482

NEWS AND VIEWS

Is the Earth alive or dead? David Lindley 483
The *ras* oncogene: A structure and some function
Irving S. Sigal 485
Structure of a ring nebula J C Raymond 486
Cretaceous unity and diversity Henry Gee 487
Topology: Mysteries of four dimensions
John D S Jones 488
Sources of increased methane G I Pearman &
P J Fraser 489
Developmental neurobiology: The milieu is the message
N Joan Abbott 490
Why Lake Michigan is not green Robert M May 491
Obituary: Sewall Wright (1889-1988)
John Maynard Smith 492
Particle physics: New phase for an old theory?
R D Peccei 492
Daedalus: Salt of the earth 493

SCIENTIFIC CORRESPONDENCE

Has the north-east Atlantic become rougher?
D J T Carter & L Draper 494
Assumptions about suicidal behaviour of aphids
M K McAllister & B D Roitberg 494
A new parameter for sex education H Sies 495
What are the masses of elementary particles? I J Good 495

BOOK REVIEWS

Who Got Einstein's Office? Eccentricity and Genius at the
Institute for Advanced Study by E Regis Daniel J Kevles 497
Seventy-five Years in Ecology: The British Ecological
Society by J Sheail Kenneth Mellanby 498
Molecules and Morphology in Evolution: Conflict or
Compromise? C Patterson ed Vincent Sarich 499
Biogeography and Plate Tectonics by J C Briggs Barry Cox
■ Cell-to-Cell Communication W C De Mello ed
Daniel Goodenough 500

ARTICLES

Response of a general circulation model to a prescribed
Antarctic ozone hole
J T Kiehl, B A Boville & B P Briegleb 501
Gas compression and jet formation in cavities collapsed by
a shock wave
J P Dear, J E Field & A J Walton 505
A specific inhibitor of the *ran1* protein kinase regulates
entry into meiosis in *Schizosaccharomyces pombe*
M McLeod & D Beach 509

Contents continued ►

Nature (ISSN 0028-0836) is published weekly on Thursday, except the last week in December, by Macmillan Magazines Ltd (4 Little Essex Street, London WC2R 3LF). Annual subscription for USA and Canada US\$250 (institutional/corporate), US\$125 (individual making personal payment). USA and Canadian orders to: *Nature*, Subscription Dept, PO Box 7663, Teaneck, NJ 07666-9857, USA. Other orders to: *Nature*, Brunel Road, Basingstoke, Hants RG21 2XS, UK. Second class postage paid at New York, NY 10012 and additional mailing offices. Authorization to photocopy material for internal or personal use, or internal or personal use of specific clients, is granted by *Nature* to libraries and others registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided the base fee of \$1.00 a copy plus \$0.10 a page is paid direct to CCC, 21 Congress Street, Salem, MA 01970, USA. Identification code for *Nature*: 0028-0836/88 \$1.00 + \$0.10. US Postmaster send address changes to: *Nature*, 65 Bleeker Street, New York, NY 10012. Published in Japan by Nature Japan K.K., Shin-Mitsuke Bldg, 36 Ichigaya Tamachi, Shinjuku-ku, Tokyo 162, Japan. © 1988 Macmillan Magazines Ltd.

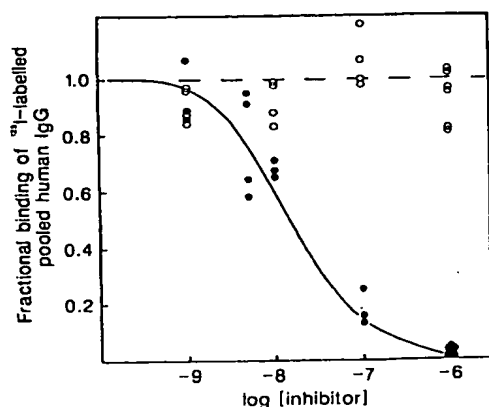


Fig. 2 Inhibition of ^{125}I -labelled pooled human IgG binding to high affinity Fc receptors (FcRI) on U937 cells by monomeric mouse IgG2b immunoglobulins. (○), Wild type IgG2b; (●), Glu 235 → Leu mutant IgG2b. For methods see Fig. 3 legend.

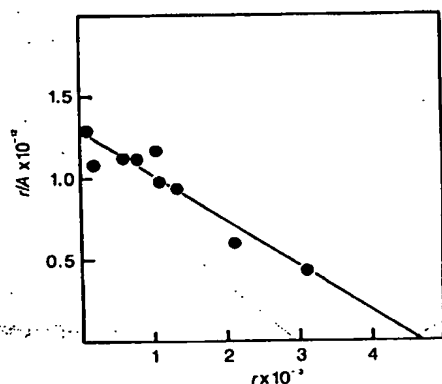


Fig. 3 Scatchard plot of ^{125}I -labelled mutant Glu235 → Leu mouse IgG2b binding to high affinity receptors (FcRI) on U937 cells. r , Number of moles of ^{125}I -(Glu 235 → Leu) mouse IgG2b antibody bound per mole of cells. A , Concentration of free ^{125}I -mutant IgG2b. The number of receptors per cell is lower than those previously reported^{14,16}, but a Scatchard analysis of ^{125}I -labelled pooled human IgG binding to the U937 cells was similar (not shown). The diminished values for receptor number may be caused by growing U937 to high cell concentrations (0.9×10^6 per ml). **Methods.** The IgG-FcRI binding assay was essentially as previously described⁸, except that after introduction of water-immiscible oil to the equilibrium mixture followed by rapid centrifugation, the pelleted cells (bound ^{125}I -IgG) and medium (free ^{125}I -IgG) were separated by slicing through the tube within the oil layer.

(cleaved between 233 and 234)¹⁸ resulted in a loss of binding to human FcRI^{19,20}, although in these two cases the two CH2 domains of the antibody are no longer tethered together by the hinge disulphides. In the alignment of ref. 12, antibodies with substitutions at residues 231 and 233 still bind tightly to FcRI, but those with changes at residue 234 have a reduced affinity. Furthermore residues 236–238 are completely conserved, except in mouse IgG1 and human IgG2, which do not bind to human FcRI. Much of the link, in particular residues 234–238, may therefore be required for binding to human FcRI.

The hinge link is mobile in the crystallographic structure of human Fc²¹ and is accessible to proteolytic attack. Thus papain cleaves between residues 233 and 234 in mouse IgG2a and IgG2b¹⁸; pepsin between residues 234 and 235 in human IgG1²² and residues 238 and 239 in mouse IgG1²³; thermolysin between residues 234 and 235 in human IgG1¹⁷. The facile proteolysis

of several IgG isotypes in this region may simply reflect the underlying design of the FcRI binding site. The site appears to be accessible and flexible and would permit, for example, a hinge dislocation on binding to FcRI²⁴.

In conclusion, our results suggest that the hinge link, either as a single flexible strand or paired with the strand from the other heavy chain, is a major determinant in binding of antibody to FcRI, and we would predict that changing Leu 235 for glutamic acid (and perhaps other side chains) would destroy the interaction of human IgG1 or IgG3 with FcRI. The possibility of turning on and off the interaction of antibody with human FcRI, could help dissect the role of this receptor in phagocytosis and cell mediated lysis and in antibody therapy. Furthermore in imaging of solid tumours, eliminating interactions with FcRI could help reduce background due to antibody binding to cells with high affinity receptors in the lymphatics, liver and spleen.

We thank M. S. Neuberger for the mouse IgG2b expression vector, M. S. Neuberger and C. Milstein for advice, and M. Clark for comments on the draft. This work was supported by the Medical Research Council and the Wellcome Trust. D.R.B. is a Jenner Fellow of the Lister Institute of Preventive Medicine.

Received 13 January; accepted 3 March 1988.

- Burton, D. R. *Molec. Immun.* 22, 161–206 (1985).
- Silverstein, S. C., Steinman, R. M. & Cohn, Z. A. *Rev. Biochem.* 46, 669–722 (1977).
- Shen, L., Guyre, P. M. & Fanger, M. W. *J. Immun.* 139, 534–538 (1987).
- Graziano, R. F. & Fanger, M. W. *J. Immun.* 139, 3536–3541 (1987).
- Karpovsky, B., Titus, J. A., Stephany, D. A. & Segal, D. M. *J. exp. Med.* 160, 1686–1701 (1984).
- Anderson, C. L. & Looney, R. J. *Immun. Today* 7, 264–266 (1987).
- Frangione, B. & Milstein, C. *Nature* 216, 939–941 (1967).
- Woolf, J. M., Nik Jaafar, M., Jefferis, R. & Burton, D. R. *Molec. Immun.* 21, 523–527 (1984).
- Leatherbarrow, R. J. *et al. Molec. Immun.* 22, 407–415 (1985).
- Partridge, L. J., Woolf, J. M., Jefferis, R. & Burton, D. R. *Molec. Immun.* 23, 1365–1372 (1986).
- Klein, M. *et al. Proc. natn. Acad. Sci. U.S.A.* 78, 524–528 (1981).
- Woolf, J. M., Partridge, L. J., Jefferis, R. & Burton, D. R. *Molec. Immun.* 23, 319–330 (1986).
- Neuberger, M. S. & Williams, G. T. *Phil. Trans. R. Soc. A317*, 425–432 (1986).
- Anderson, C. L. & Abraham, G. N. *J. Immun.* 125, 2735–2741 (1980).
- Kurlander, R. J. & Barker, J. *J. clin. Invest.* 69, 1–8 (1982).
- Fries, L. F., Hall, R. P., Lawley, T. J., Crabtree, G. R. & Frank, M. M. *J. Immun.* 129, 1041–1049 (1982).
- Hunneyball, I. M. & Stanworth, D. R. *Immunology* 30, 579–586 (1976).
- Frankus, T. & Birshstein, B. K. *Biochemistry* 17, 4324–4331 (1978).
- Ratcliffe, A. & Stanworth, D. R. *Immunology* 50, 93–100 (1983).
- McCool, D., Birshstein, B. K. & Palmer, R. H. *J. Immun.* 135, 1975–1980 (1985).
- Deisenhofer, J. *Biochemistry* 20, 2361–2370 (1981).
- Frangione, B. & Milstein, C. *J. molec. Biol.* 33, 893–906 (1968).
- Svavil, J. & Milstein, C. *Nature* 228, 930–935 (1970).
- Burton, D. R. *Immun. Today* 7, 165–167 (1986).
- Carter, P., Bedouelle, H. & Winter, G. *Nucleic Acids Res.* 13, 4431–4443 (1985).
- Raychaudhuri, G., McCool, D. & Panter, R. H. *Molec. Immun.* 22, 1009–1019 (1985).

Dissecting the catalytic triad of a serine protease

Paul Carter & James A. Wells

Department of Biomolecular Chemistry, Genentech Inc.,
460 Point San Bruno Boulevard, South San Francisco,
California 94080, USA

Serine proteases are present in virtually all organisms and function both inside and outside the cell¹; they exist as two families, the 'trypsin-like' and the 'subtilisin-like', that have independently evolved a similar catalytic device² characterized by the Ser, His, Asp triad, an oxyanion binding site, and possibly other determinants that stabilize the transition state (Fig. 1)^{3–4}. For *Bacillus amyloliquefaciens* subtilisin, these functional elements impart a total rate enhancement of at least 10^9 to 10^{10} times the non-enzymatic hydrolysis of amide bonds. We have examined the catalytic importance and interplay between residues within the catalytic triad by individual or multiple replacement with alanine(s), using site-directed mutagenesis^{5,6} of the cloned *B. amyloliquefaciens* subtilisin gene⁷. Alanine substitutions were chosen to minimize unfavourable steric contacts and to avoid imposing new charge interactions or hydrogen bonds from

Table 1 Kinetic parameters of mutant subtilisins with the substrate *N*-succinyl-L-Ala-L-Ala-L-Pro-L-Phe-*p*-nitroanilide at pH 8.60

Enzyme	Active site configuration			k_{cat} (s^{-1})	K_m (μM)	k_{cat}/K_m ($s^{-1} M^{-1}$)	$k_{cat}(\text{mutant})/k_{cat}(\text{S24C})$
	Ser221	His64	Asp32				
Wild type	+	+	+	$(4.4 \pm 0.1) \times 10^1$	180 ± 10	$(2.5 \pm 0.1) \times 10^5$	0.74 ± 0.01
S24C	+	+	+	$(5.9 \pm 0.2) \times 10^1$	220 ± 20	$(2.7 \pm 0.2) \times 10^5$	1
S24C:S221A	-	+	+	$(3.4 \pm 0.1) \times 10^{-5}$	420 ± 40	$(8.2 \pm 0.6) \times 10^{-2}$	$(5.8 \pm 0.1) \times 10^{-7}$
S24C:H64A	+	-	+	$(3.8 \pm 0.2) \times 10^{-5}$	390 ± 50	$(9.6 \pm 1.0) \times 10^{-2}$	$(6.4 \pm 0.2) \times 10^{-7}$
S24C:D32A	+	+	-	$(2.3 \pm 0.2) \times 10^{-3}$	480 ± 80	4.7 ± 0.7	$(3.8 \pm 0.2) \times 10^{-5}$
S24C:D32A:H64A	+	-	-	$(2.6 \pm 0.1) \times 10^{-4}$	270 ± 50	$(9.4 \pm 1.6) \times 10^{-1}$	$(4.3 \pm 0.1) \times 10^{-6}$
S24C:H64A:S221A	-	-	+	$(2.8 \pm 0.2) \times 10^{-5}$	290 ± 40	$(9.6 \pm 1.3) \times 10^{-2}$	$(4.8 \pm 0.2) \times 10^{-7}$
S24C:D32A:S221A	-	+	-	$(2.8 \pm 0.1) \times 10^{-5}$	310 ± 40	$(9.2 \pm 0.9) \times 10^{-2}$	$(4.8 \pm 0.1) \times 10^{-7}$
S24C:D32A:H64A:S221A	-	-	-	$(3.0 \pm 0.1) \times 10^{-5}$	230 ± 20	$(1.3 \pm 0.1) \times 10^{-1}$	$(5.1 \pm 0.1) \times 10^{-7}$
				k_{buffer} (s^{-1})			
No enzyme	none			$(1.1 \pm 0.1) \times 10^{-8}$	-	-	$(1.9 \pm 0.1) \times 10^{-10}$

Mutants are abbreviated by the single-letter code for the wild-type amino acid followed by its codon position and the amino acid replacement; multiple mutants are designated by listing single mutant components separated by colons (for example, double mutant Ser24 to Cys, Ser221 to Ala is designated S24C:S221A). Construction of the mutants S24C and H64A and the double mutant S24C:H64A was as described^{12,13}. The mutations D32A and S24C were constructed simultaneously using a 48-mer oligonucleotide^{5,6} and the S221A mutant was constructed by cassette mutagenesis²⁵. The remaining multiple mutants were constructed by 3-way ligations using a 6 kb *EcoRI*/*Bam*HI fragment from the vector pSS5 (B. Cunningham, D. Powers, and J. W. unpublished) and two subtilisin fragments from appropriate mutants. Mutant constructions were verified by dideoxy sequencing²⁶. Mutant plasmids were expressed in a protease deficient strain of *B. subtilis*, BG2036²⁷. Rescue of active site mutants by co-culturing with the mutant A48E and purification was as described¹². Mutant subtilisins were assayed with the substrate, *N*-succinyl-L-Ala-L-Ala-L-Pro-L-Phe-*p*-nitroanilide (Sigma). Six hydrolysis assays were performed simultaneously against substrate blanks in 1 ml 100 mM Tris-HCl (pH 8.60) 4% (v/v) dimethylsulphoxide at $(25 \pm 0.2)^\circ C$ using a Kontron Uvikon 860 spectrophotometer. Initial reaction rates were determined from the increase in absorbance at 410 nm on release of *p*-nitroaniline ($\epsilon_{410} = 8,480 M^{-1} cm^{-1}$)²⁸. The total substrate concentration in each assay was determined from the A_{410} after complete hydrolysis. The initial rate data were fitted to the Michaelis-Menten relationship using least squares analysis to determine K_m and V_{max} . Turnover number (k_{cat}) was calculated from the spectrophotometrically determined enzyme concentration ($\epsilon_{280}^{0.1\%} = 1.17$)²⁹. Enzyme concentrations in the assays were 30–110 $\mu g ml^{-1}$ for the active site mutants and 1 $\mu g ml^{-1}$ for the wild type and S24C enzymes. Catalytic triad residues are represented by (+) and Ala replacements by (-). Data are presented \pm standard errors and the spontaneous hydrolysis rate of substrate under these conditions is shown as k_{buffer} .

substituted side chains. In contrast to the effect of mutations in residues involved in substrate binding^{8–10}, the mutations in the catalytic triad greatly reduce the turnover number and cause only minor effects on the Michaelis constant. Kinetic analyses of the multiple mutants demonstrate that the residues within the triad interact synergistically to accelerate amide bond hydrolysis by a factor of $\sim 2 \times 10^6$.

Subtilisin is synthesized as a membrane-associated precursor (preprosubtilisin)⁷. When expressed in a protease-deficient strain of *B. subtilis*, mature *B. amyloliquefaciens* subtilisin is efficiently released into the medium after autoproteolytic cleavage¹¹. Mutagenesis of the catalytic residues in subtilisin (which essentially inactivates the protease) disrupts this processing, but processing can be restored by co-culturing the mutants with a small amount of a *B. subtilis* strain (called a 'helper') harbouring an active subtilisin gene¹². We have constructed a series of active site mutants in which the catalytic triad residues are replaced by alanine in every possible combination (ref. 12, Table 1). Each mutant also contains a surface-accessible Ser24 to Cys mutation¹³ designated S24C (mutant enzymes are named using the single letter code for amino acids to indicate the substitutions made, see Table 1). The S24C substitution permits reversible attachment to an activated thiol sepharose column thereby eliminating traces of contaminating helper subtilisin which is cysteine-free¹².

The hydrolysis of the substrate (*N*-succinyl-L-Ala-L-Ala-L-Pro-L-Phe-*p*-nitroanilide) by most of the active site mutants produced only small absorbance changes (ΔA_{410} of 0.01 to 0.10) over long periods (up to 12 h), yet the data exhibit typical Michaelis-Menten saturation behaviour (Fig. 2) with standard errors almost as small as those for wild-type subtilisin (Table 1). No detectable loss of catalytic activity occurred even during the longest kinetic runs. In addition, the background (non-enzymatic) hydrolysis of substrate was $\leq 25\%$ of the catalysed rate for even the least active enzymes (Fig. 2). The non-enzymatic

hydrolysis was subtracted directly from the enzyme assays using blank substrate solutions in a double beam spectrophotometer.

Kinetic analysis of the active site single mutants (Table 1) shows that replacement of the catalytic serine, histidine or aspartate causes a drop in turnover number (k_{cat}) by factors of 2×10^6 , 2×10^6 and 3×10^4 , respectively. The 100-fold lower values of k_{cat} which result from substitution of Ser221 and His64, compared with Asp32, are consistent with their more central role in catalysis (Fig. 1). Each mutation causes a small increase in the Michaelis constant (K_m) (~ 2 -fold) which may result from slightly altered substrate binding contacts. (Wild-type subtilisin has a two-step enzyme mechanism where deacylation is > 33 times faster than acylation¹⁴, so that K_m is a good approximation of the enzyme-substrate dissociation constant (K_s)¹⁵. As the enzyme mechanism must be changed for at least some of the mutants (see below), K_m may be less than K_s .)

Additional mutagenesis of the S24C:S221A enzyme to replace either Asp32, His64 or both, causes essentially no further change in k_{cat} or K_m (Table 1). By comparison, further mutagenesis of the S24C:D32A parent enzyme to substitute His64 or both His64 and Ser221, further reduces k_{cat} by 9 and 76-fold, respectively, with essentially no change in K_m . These data suggest that His64 provides a catalytic advantage of ~ 10 -fold to the S24C:D32A enzyme, and that Ser221 provides ~ 10 -fold advantage to the S24C:D32A:H64A enzyme. As with the S24C:S221A family of mutants, additional mutations in the S24C:H64A enzyme to replace Ser221 or both Ser221 and Asp32 do not affect k_{cat} . But replacement of Asp32 alone in the S24C:H64A mutant to give S24C:D32A:H64A, actually increases k_{cat} 7-fold. Thus, Asp32 is a liability to the S24C:H64A enzyme, possibly because of an unfavourable electrostatic effect upon catalysis (see below).

The single and multiple mutant analyses show that the catalytic effects are non-additive in two ways. First, there is a gross discrepancy between the relative drop in k_{cat} resulting from the triple alanine mutant (2×10^6 , Table 1) compared with

Table 2 Kinetic parameters of mutant subtilisins with the substrate *N*-succinyl-L-Ala-L-Ala-L-Pro-L-Phe-*p*-nitroanilide at pH 9.70

Enzyme	Active site configuration			k_{cat} (s^{-1})	K_m (μM)	k_{cat}/K_m ($s^{-1} M^{-1}$)	k_{cat} (pH 9.7)
	Ser221	His64	Asp32				k_{cat} (pH 8.6)
Wild type	+	+	+	$(6.3 \pm 0.1) \times 10^4$	440 ± 30	$(1.4 \pm 0.1) \times 10^5$	1.4 ± 0.1
S24C	+	+	+	$(8.1 \pm 0.2) \times 10^4$	560 ± 30	$(1.5 \pm 0.1) \times 10^5$	1.4 ± 0.1
S24C:S221A	-	+	+	$(5.4 \pm 0.3) \times 10^{-5}$	650 ± 90	$(8.4 \pm 1.0) \times 10^{-2}$	1.6 ± 0.1
S24C:H64A	+	-	+	$(1.9 \pm 0.1) \times 10^{-4}$	1300 ± 150	$(1.5 \pm 0.2) \times 10^{-1}$	5.1 ± 0.2
S24C:D32A	+	+	-	$(1.8 \pm 0.1) \times 10^{-2}$	1400 ± 120	$(1.3 \pm 0.1) \times 10^1$	7.8 ± 0.4
S24C:D32A:H64A	+	-	-	$(1.8 \pm 0.1) \times 10^{-3}$	460 ± 40	3.8 ± 0.3	6.9 ± 0.3
S24C:H64A:S221A	-	-	+	$(5.2 \pm 0.2) \times 10^{-5}$	480 ± 60	$(1.1 \pm 0.1) \times 10^{-1}$	1.9 ± 0.1
S24C:D32A:S221A	-	+	-	$(5.9 \pm 0.3) \times 10^{-5}$	460 ± 80	$(1.3 \pm 0.2) \times 10^{-1}$	2.1 ± 0.1
S24C:D32A:H64A:S221A	-	-	-	$(7.8 \pm 0.3) \times 10^{-5}$	730 ± 70	$(1.1 \pm 0.1) \times 10^{-1}$	2.6 ± 0.1
				$k_{buffer} (s^{-1})$			
No enzyme	none			$(2.8 \pm 0.1) \times 10^{-8}$	-	-	2.5 ± 0.1

Kinetic data were determined as for Table 1 except that 100 mM 3-[cyclohexylamino]-2-hydroxyl-1-propane buffer (pH 9.70) was used. Ionic strength was normalized with NaCl.

the product of the relative effects from the three single alanine mutants ($\sim 10^{17}$). Second, the double alanine mutants that retain singly the catalytic Ser, His or Asp are only a factor of 8, 0.9 or 0.9 larger in k_{cat} , respectively, than the triple alanine mutant. The product of these values (~ 6) is much below the relative k_{cat} value of 2×10^6 for wild type (S24C) compared with the triple alanine mutant. Thus, non-additive effects are shown either by subtraction of catalytic residues relative to wild-type enzyme or by addition of single catalytic residues relative to the triple alanine mutant.

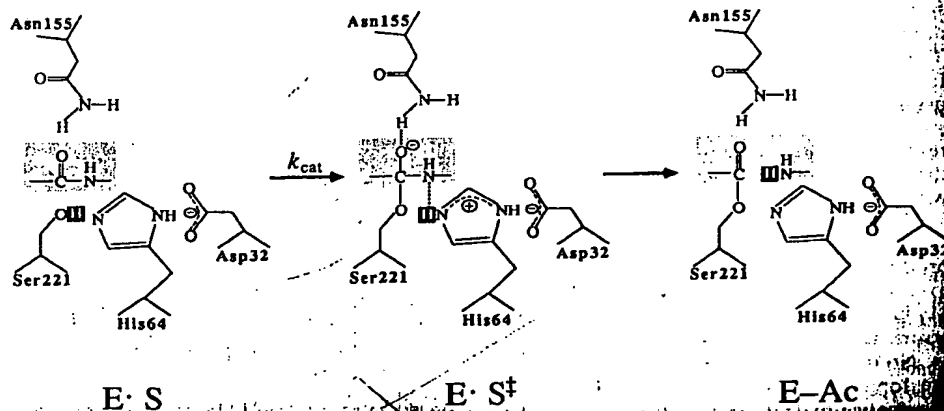
Replacement of residues in the catalytic triad with alanines necessarily perturbs the enzyme mechanism. In particular, it has been observed that in the absence of the catalytic His64 in subtilisin¹² or the catalytic Asp102 in trypsin^{16,17}, there is a marked increase in the hydroxide dependence of catalysis between pH 8 and 10 compared to the wild-type enzymes. Comparisons of the kinetic parameters for all of the catalytic triad mutants at pH 9.70 and pH 8.60 (Table 2) show that those retaining Ser221 have a substantially stronger pH dependence of k_{cat} (increased 5- to 8-fold) than enzymes containing an intact catalytic triad (increased 1.4-fold), or enzymes lacking Ser221 (increased 1.6- to 2.6-fold), or when compared with the non-enzymatic rate (increased 2.5-fold). For all enzymes the K_m values at pH 9.70 are increased between 1.5 and 3.3-fold. Preliminary evidence suggests that this effect upon K_m may result (at least partially) from ionization of Tyr104, resulting in electrostatic repulsion of the P5 succinyl group (see Fig. 3, and D.

Estell, T. Graycar, D. Powers and J. A. Wells, unpublished results).

For mutants that retain Ser221, the simplest interpretation of the data is that they continue to use Ser221 as the catalytic nucleophile. The presence of Ser221 provides a catalytic advantage of ~ 10 -fold to the S24C:D32A:H64A enzyme and ~ 100 -fold to the S24C:D32A enzyme. Furthermore, replacing His64 in the S24C:D32A enzyme causes k_{cat} to drop ~ 10 -fold, suggesting that His64 functions here to some extent (presumably as a proton acceptor for the nucleophilic Ser221). In addition, if deprotonation of the Ser221 hydroxyl is a prerequisite for nucleophilic attack in these mutants, then it is reasonable for k_{cat} to depend on hydroxide ion concentration, as observed (Table 2). Finally, in the absence of His64, the catalytic aspartate should inhibit deprotonation of Ser221 and have a deleterious electrostatic effect upon k_{cat} , as indeed was found (k_{cat} for S24C:H64A is 10-fold lower than the k_{cat} for S24C:D32A:H64A in Table 1). Like wild-type subtilisin, we anticipate the S221A family of enzymes should have a two-step enzyme mechanism. For these mutants, if deacylation is rate-determining, it is possible that the K_m values are substantially less than the K_d values¹⁵.

For the S24C:S221A family of enzymes, the reaction cannot proceed by the usual serine acyl-enzyme intermediate. Instead, direct attack of water on the scissile peptide bond may occur to produce a single tetrahedral intermediate that collapses to give the hydrolysed products. Nucleophilic attack by water is

Fig. 1 Schematic diagram showing the rate limiting acylation step in the hydrolysis of peptide bonds by subtilisin. In going from the Michaelis enzyme-substrate complex ($E \cdot S$) to the transition state complex ($E \cdot S^\ddagger$), the proton on Ser221 (darkly shaded) is transferred to His 64, thus permitting nucleophilic attack on the scissile peptide bond²⁻⁴. The proton is then transferred to the amine leaving group to generate the acyl-enzyme intermediate ($E-Ac$). Asp32 (as for Asp102 in trypsin^{14,16,17}) is believed to position the correct tautomer of His64 for catalysis in the $E \cdot S$ complex and stabilize the protonated form of His64 in the $E \cdot S^\ddagger$ complex. Some of the hydrogen bonds that form in the $E \cdot S^\ddagger$ complex are shown by dotted lines. In deacylation these steps are reversed and water (as the nucleophile) replaces the amine leaving group.



consistent with the weak hydroxide dependence of k_{cat} for the S24C:S221A-containing mutants. The lack of a deleterious electrostatic effect from Asp32 is also consistent with a neutral attacking nucleophile (compare S24C:H64A:S221A with S24C:D32A:H64A:S221A in Table 1). It is unlikely that the S221A group of enzymes use the other members of the catalytic triad because there is no additional kinetic advantage for including the His64 or Asp32. (Strictly, we cannot be sure that the residual members of the triad are catalytically inert. We simply cannot detect any catalytic advantage for them over the residual activity resulting from determinants unrelated to the triad—see below). Preliminary X-ray analysis of the S221A enzyme indicates no large structural change except for the Ser221 to Ala substitution (R. Bott and M. Ultsch, personal communication). More kinetic and structural data will be necessary however, to substantiate the possible mechanisms discussed above.

The small values of k_{cat} for the active site mutants raise questions regarding protease contaminants or assay artefacts. The following evidence argues strongly against these possibilities. (1) Unlike wild-type subtilisin, the mutant enzymes are not inhibited by phenylmethylsulphonyl fluoride. (2) Although changes in the K_m values are small for these mutants, many are statistically different from wild type (Tables 1, 2). A contamination with helper subtilisin (regardless of amount) would give a constant value for the K_m equal to wild type. (3) Many of the active site mutants differ significantly from each other in k_{cat} and K_m at pH 8.6 (Table 1), which is inconsistent with a constant contaminant. (4) The mutants differ among themselves and wild type in terms of their pH dependence of k_{cat} (Table 2), a result inconsistent with a fixed protease contaminant. (5) Although the kinetic values reported in Tables 1 and 2 are from the same batch of enzyme, most mutant enzymes have been purified more than once. In every repeat case (data not shown) the kinetic values agree within the standard error limits shown ($\pm 15\%$ for k_{cat} and K_m), even though enzyme yields varied, and purification protocols were sometimes slightly modified. (6) The mutants were expressed in an extracellular protease deficient strain of *B. subtilis*, purified on activated thiol sepharose, and judged to be >99% pure by silver-stained SDS-PAGE. Moreover, further purification of the S24C:H64A enzyme by native gel electrophoresis gave identical kinetic values as the starting material¹².

It is formally possible that the residual activity in some or all of these mutants occurs at a non-specific site(s) distinct from the active site. The following points argue for catalysis at the active site. (1) In some cases the kinetic effects are cumulative for mutagenesis at the active site. For example, the k_{cat} values decrease in the following order: S24C>S24C:D32A>S24C:D32A:H64A>S24C:D32A:H64A:S221A (Table 1). (2) The K_m values are usually not more than twofold above the wild type value suggesting continued strong and specific binding (assuming $K_m \sim K_s$). Furthermore, the active site mutants show a similar pH dependent increase in K_m as wild type subtilisin. (3) The substrate preferences for the S24C:D32A and S24C:S221A enzymes toward two other substrates essentially parallel the wild type enzyme (P. C., unpublished results). The substrate specificity of the S24C:H64A enzyme also parallels the wild type except for a strong preference for His P2 substrates¹² (see below). (4) The activity of the S24C:H64A enzyme is heat denaturable (C. Mitchinson, unpublished results) which indicates that the native protein conformation is critical for catalysis. (5) The residual activity for even the least active mutant is still > 10^3 fold above the non-enzymatic rate. This catalytic rate is in the range measured for 'good' catalytic antibodies^{18,20}. Taken together these data provide compelling evidence that the residual catalytic activities we have measured are not due to protease contamination, assay artefacts or non-specific catalysis away from the normal active site.

We suggest that the residual activity in the triple mutant is derived from remaining binding determinants which stabilize

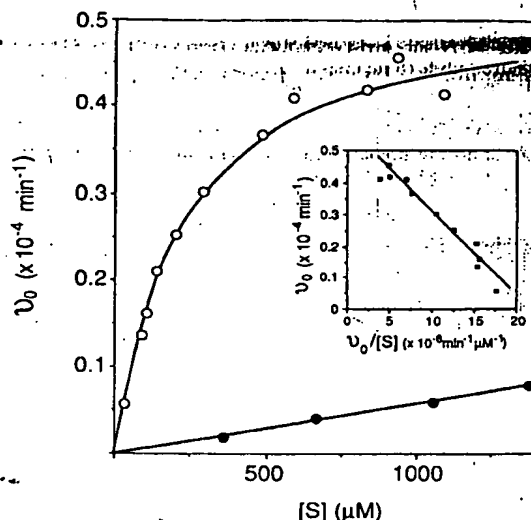
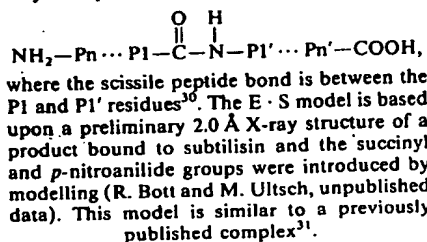


Fig. 2 Initial rate of hydrolysis v_0 ($\Delta A_{410}/\Delta t$) versus the concentration of the substrate *N*-succinyl-L-Ala-L-Ala-L-Pro-L-Phe-*p*-nitroanilide [S] in the absence (●) or presence (○) of S24C:D32A:H64A:S221A subtilisin. The background hydrolysis rate (●) was subtracted directly from the rate in the presence of subtilisin to give the enzymatic rate (○). Experiments were performed in 100 mM Tris · HCl, pH 8.60, at $25 \pm 0.2^\circ\text{C}$, as described in Table 1. Insert (■) shows an Eadie-Hofstee plot of the initial rate data.

the transition state complex outside the catalytic triad. In fact, previous data show that when the hydrogen bond to Asn155 in the oxyanion binding site (Fig. 1) is disrupted by site-directed mutagenesis, there is a 10^2 to 10^3 drop in k_{cat} with little effect upon K_m ^{14,21}. Additional hydrophobic interactions (Fig. 3) with the P1 substrate side chain⁸ and binding interactions with the P2 to P4 substrate residues^{22,23} are estimated to contribute independently factors of 10 to 100 to k_{cat} . Structural analysis²⁴ suggests there are additional hydrogen bonds in the transition state complex between the NH of Ser221 and the oxyanion, and between the NH of the P1 substrate residue and the carbonyl of Ser125. Deriving the total catalytic contribution from the sum of these individual binding components may lead to overestimation because of their possible interdependence. Nonetheless, our data indicate that some or all of these determinants are important for stabilizing the tetrahedral transition state complex (contributing > 10^3 to k_{cat}), and are not simply required for positioning the substrate for optimal nucleophilic attack by Ser221.

From an evolutionary point of view, it is extremely unlikely that the catalytic triad arose in one step rather than involving active intermediates. This view is now apparently complicated by the fact that the residues in the catalytic triad function in an extremely synergistic manner. But, assuming that the present-day enzyme is a reasonable model of its ancestor, there are at least two possible mutagenic pathways that give progressive increases in catalytic rate by stepwise introduction of the residues in the triad. In the first pathway, installing Ser221 followed by His64 and then Asp32 gives progressive increases of 8, 9 and 3×10^4 in k_{cat} (Table 1). This progression is even more uniform under alkaline conditions, resulting in increases in k_{cat} of 50, 10 and 5×10^3 (Table 2). A second mutagenic pathway is possible by preferential use of a His P2 substrate (Fig. 3)³⁰ in place of the catalytic His64. We have previously shown that the Ala64 enzyme has a turnover number of $2 \times 10^{-2} \text{ s}^{-1}$ for hydrolysis of a His P2 substrate compared to $8 \times 10^{-6} \text{ s}^{-1}$ for an Ala P2 substrate¹². This catalytic advantage, which we have called 'substrate-assisted catalysis', makes it feasible to reverse the order of introducing His64 and Asp32.

Fig. 3 Stereoview of a model containing the substrate, *N*-succinyl-L-Ala-L-Ala-L-Pro-L-Phe-*p*-nitroanilide (bold lines and filled atoms), bound to the active site of *B. amyloliquefaciens* subtilisin. Alpha carbons from important enzyme and substrate residues are labelled. In protease substrate nomenclature the substrate may be represented as



Of course this advantage would apply only to His P2 substrates but would be reasonable if the ancestral enzyme were involved in specific proteolytic processing, for example. Regardless of the exact order of evolutionary events, our mutagenic studies show that inserting catalytic triad residues in a stepwise fashion can produce enzyme intermediates with progressively increased turnover numbers.

In summary, when residues in the catalytic triad are altered separately or together there are large effects on turnover rate, consequent changes in the enzyme mechanism, and only minor effects on the Michaelis constant. The residues in the catalytic triad function in a strongly synergistic fashion and contribute a factor of about 2×10^6 to the total to the catalytic rate enhance-

ment of 10^9 to 10^{10} . The residual activity from complete replacement of the catalytic triad is not a contaminant or other artefact, but results from transition state stabilization from contacts outside the catalytic triad. Finally, despite the synergy between the catalytic triad residues, their sequential introduction is reasonable in terms of both evolution and function.

We thank Dr Rick Bott for help in preparing Fig. 3 and sharing unpublished X-ray coordinates, Dr Polly Moore and Ann-Benninger for assistance in data handling, the organic chemistry group at Genentech for synthesis of oligonucleotides, and Drs Tony Kossiakoff, Jack Kirsch and Ron Wetzel for helpful comments on this manuscript.

Received 19 January 1988; accepted 22 February 1988.

1. Stroud, R. M. *Science* **181**, 74-88 (1974).
2. Kraut, J. A. *Rev. Biochem.* **46**, 331-358 (1977).
3. Fink, A. L. in *Enzyme Mechanisms* (eds Page, M. I. & Williams, A.) 159-177 (Roy. Soc. Chem. 1987).
4. Kossiakoff, A. A. in *Biological Macromolecules and Assemblies* Vol. 3 (eds Jurnak, F. A. & McPherson, A.) 370-412 (1987).
5. Zoller, M. J. & Smith, M. *Nucleic Acids Res.* **10**, 6487-6500 (1982).
6. Carter, P., Bedouelle, H. & Winter, G. *Nucleic Acids Res.* **13**, 4431-4443 (1986).
7. Wells, J. A., Ferrari, E., Henner, D. J., Estell, D. A. & Chen, E. Y. *Nucleic Acids Res.* **11**, 7911-7925 (1983).
8. Estell, D. A. *et al. Science* **233**, 659-663 (1986).
9. Wells, J. A., Powers, D. B., Bott, R. R., Graycar, T. P. & Estell, D. A. *Proc. natn. Acad. Sci. U.S.A.* **84**, 1219-1223 (1987).
10. Wells, J. A., Cunningham, B. C., Graycar, T. P. & Estell, D. A. *Proc. natn. Acad. Sci. U.S.A.* **84**, 5167-5171 (1987).
11. Power, S. D., Adams, R. M. & Wells, J. A. *Proc. natn. Acad. Sci. U.S.A.* **83**, 3096-3100 (1986).
12. Carter, P. & Wells, J. A. *Science* **237**, 394-399 (1987).
13. Wells, J. A. & Powers, D. B. *J. biol. Chem.* **261**, 6564-6570 (1986).
14. Wells, J. A., Cunningham, B. C., Graycar, T. P. & Estell, D. A. *Phil. Trans. R. Soc. A* **317**, 415-423 (1986).

15. Gutfreund, H. & Sturtevant, J. M. *Biochem. J.* **63**, 656-661 (1956).
16. Craik, C. S., Rocznik, S., Largman, C. & Rutter, W. J. *Science* **237**, 909-913 (1987).
17. Sprang, S. *et al. Science* **237**, 905-909 (1987).
18. Tramontano, A., Janda, K. D. & Lerner, R. A. *Science* **234**, 1566-1570 (1986).
19. Pollack, S. J., Jacobs, J. W. & Schultz, P. G. *Science* **234**, 1570-1573 (1986).
20. Napper, A. D., Benkovic, S. J., Tramontano, A. & Lerner, R. A. *Science* **237**, 1041-1043 (1987).
21. Bryan, P., Pantoliano, M. W., Quill, S. G., Hsiao, H.-Y. & Poulos, T. *Proc. natn. Acad. Sci. U.S.A.* **83**, 3743-3745 (1986).
22. Morihara, K., Oka, T. & Tsuzuki, H. *Arch. Biochem. Biophys.* **138**, 515-525 (1970).
23. Morihara, K., Oka, T. & Tsuzuki, H. *Biochem. biophys. Res. Commun.* **35**, 210-214 (1969).
24. Robertus, J. D., Kraut, J., Alden, R. A. & Birktoft, J. J. *Biochemistry* **11**, 4293-4303 (1972).
25. Wells, J. A., Vasser, M. & Powers, D. B. *Gene* **34**, 315-323 (1985).
26. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463-5467 (1977).
27. Yang, M. Y., Ferrari, E. & Henner, D. J. *J. Bact.* **160**, 15-21 (1984).
28. DelMar, E. G., Langman, C., Brodrick, J. W. & Goekas, M. C. *Analyt. Biochem.* **99**, 316-320 (1979).
29. Matubara, H., Kasper, C. B., Brown, D. M. & Smith, E. L. *J. biol. Chem.* **240**, 1125-1130 (1965).
30. Schechter, I. & Berger, A. *Biochem. biophys. Res. Commun.* **27**, 157-162 (1967).
31. Robertus, J. D. *et al. Biochemistry* **11**, 2439-2449 (1972).

Erratum Oxygen isotope dating of the Australian regolith

Michael I. Bird & Allan R. Chivas
Nature **331**, 513-516 (1988)

In this letter, Fig. 1 as printed is too small to allow the symbols to be properly differentiated. The figure is reprinted here with enlarged symbols. In addition, line 13 in the left-hand column on page 515 has become garbled by an error in a line correction. The first sentence in that paragraph should read: "Isotopic results obtained from residual clays (collected *in situ* from regolith profiles) of post-mid-Tertiary age have $\delta^{18}\text{O}$ values between 17.5 and 21.3‰, with the exception of samples from field 'd' (representing latitudes north of $\sim 20^\circ\text{S}$) which have anomalously low values."

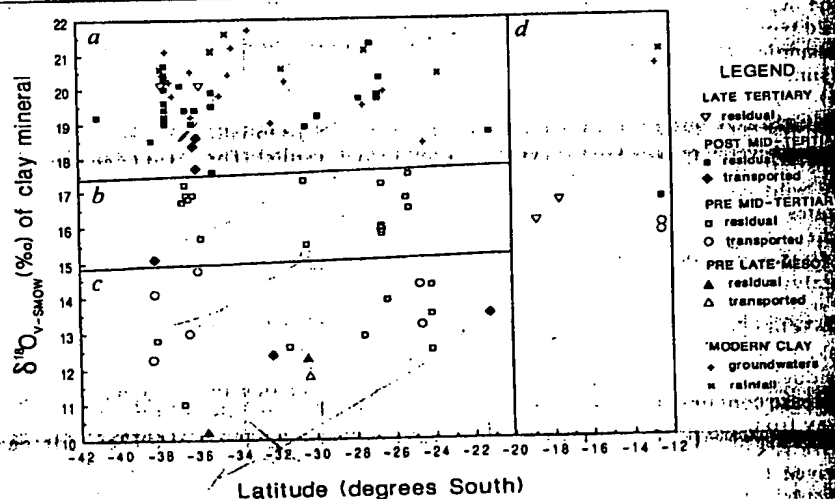


Exhibit 9

Site-directed Mutagenesis Suggests Close Functional Relationship between a Human Rhinovirus 3C Cysteine Protease and Cellular Trypsin-like Serine Proteases*

(Received for publication, November 13, 1989)

Keat-Chye Cheah†, Louis E.-C. Leong, and Alan G. Porter§

From the Institute of Molecular and Cell Biology, National University of Singapore, Kent Ridge Crescent, Singapore 0511

Human rhinoviruses, like other picornaviruses, encode a cysteine protease (designated 3C) which cleaves mainly at viral Gln-Gly pairs. There are significant areas of homology between picornavirus 3C cysteine proteases and cellular serine proteases (e.g. trypsin), suggesting a functional relationship between their catalytic regions. To test this functional relationship, we made single substitutions in human rhinovirus type 14 protease 3C at seven amino acid positions which are highly conserved in the 3C proteases of animal picornaviruses. Substitutions at either His-40, Asp-85, or Cys-146, equivalent to the trypsin catalytic triad His-57, Asp-102, and Ser-195, respectively, completely abolished 3C proteolytic activity. Single substitutions were also made at either Thr-141, Gly-158, His-160, or Gly-162, which are equivalent to the trypsin specificity pocket region. Only the mutant with a conservative Thr-141 to Ser substitution exhibited proteolytic activity, which was much reduced compared with the parent. These results, together with immunoprecipitation data which indicate that Asp-85, Thr-141, and Cys-146 lie in accessible surface regions, suggest that the catalytic mechanism of picornavirus 3C cysteine proteases is closely related to that of cellular trypsin-like serine proteases.

Human rhinoviruses (HRVs),¹ the main causative agents of the common cold, form one genus of the Picornavirus family (Stott and Killington, 1972; Gwaltney, 1975). The primary translation product of the positive stranded RNA genome of picornaviruses (e.g. HRVs, poliovirus, and foot-and-mouth disease virus) is a single precursor polypeptide which is rapidly processed by viral proteases to mature products (Nicklin *et al.*, 1986; Kräusslich and Wimmer, 1988). Proteolytic cleavage of the viral precursor protein plays an important part in the regulation of picornavirus replication. Two Tyr-Gly pairs in the precursor are cleaved by viral protease 2A (Kräusslich and Wimmer, 1988). Most of the cleavages are performed by viral protease 3C (3C^{pro}) which

exhibits a preference for Gln-Gly pairs (Nicklin *et al.*, 1986; Kräusslich and Wimmer, 1988).

3C^{pro} from poliovirus (Hanecak *et al.*, 1984; Ivanoff *et al.*, 1986; Richards *et al.*, 1987; Nicklin *et al.*, 1988), encephalomyocarditis virus (Parks *et al.*, 1989), foot-and-mouth disease virus (Klump *et al.*, 1984; Strebel *et al.*, 1986) and HRV-14 (Cheah *et al.*, 1988; Libby *et al.*, 1988) have been cloned and expressed in *Escherichia coli*. In most of these studies, the 3C^{pro} precursor form has been shown to cleave its flanking Gln-Gly sites to release mature 3C^{pro} in an autocatalytic fashion. However, cleavage at Gln-Gly to release the poliovirus capsid proteins is performed not by 3C^{pro} but by the 3C-3D precursor in which 3C^{pro} is covalently fused to the adjacent 3D polymerase (Jore *et al.*, 1988; Ypma-Wong *et al.*, 1988).

3C^{pro} activity is inhibited by cysteine protease inhibitors, indicating that cysteine may be an active-site amino acid (Korant, 1973; Pelham, 1978; Korant *et al.*, 1985). In fact, sequence comparisons of 3C proteases from animal picornaviruses and 3C-like proteases from some plant viruses showed that only one of the cysteines (Cys-147 in poliovirus) is highly conserved in all these viruses (Argos *et al.*, 1984; Franssen *et al.*, 1984). Strong evidence that Cys-147 of poliovirus is an active-site amino acid came from site-directed mutagenesis studies which demonstrated that mutation of the highly conserved Cys-147 to Ser resulted in the inactivation of the protease, whereas similar mutation of the nonconserved Cys-153 had no effect (Ivanoff *et al.*, 1986).

It was suggested on the basis of computer alignments that the viral 3C cysteine proteases may represent an evolutionary link between the cellular cysteine proteases exemplified by papain, and the cellular trypsin-like serine proteases (Gorbalenya *et al.*, 1986). More extensive computer alignment of picornavirus 3C proteases and cellular serine proteases revealed some remarkable primary and secondary structural homologies, indicating that certain amino acids within 3C^{pro}, including Cys-147 (Cys-146 in HRV-14), may be responsible for catalysis or substrate binding in a mechanistically similar fashion to the cellular serine proteases (Bazan and Fletterick, 1988). His-40, Asp-85, and Cys-146 of HRV-14 3C^{pro}, which are completely conserved in all picornaviruses align with His-57, Asp-102, and Ser-195 of the trypsin-like serine protease catalytic triad (Bazan and Fletterick, 1988). As a result of these alignments, Thr-141, Gly-158, and His-160 of HRV-14 3C^{pro} which are also completely conserved in all picornaviruses, and Gly-162 which is conserved in HRVs and enteroviruses (e.g. poliovirus), align with the amino acids lying in or close to the specificity pocket of the cellular serine proteases (Bazan and Fletterick, 1988). In this paper, we describe introduction of single amino acid substitutions in HRV-14 3C^{pro} at the positions which correspond to the trypsin catalytic triad and specificity pocket. All except one of the substitutions

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom all correspondence should be sent.

† Present address: Dept. of Microbiology and Immunology, The University of Adelaide, Box 498, GPO, Adelaide, South Australia 5001, Australia.

¹ The abbreviations used are: HRVs, human rhinoviruses; HRV-14, human rhinovirus type 14; 3C^{pro}, viral protease 3C; SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; KLH, key-hole limpet hemocyanin; Ap^R, ampicillin resistant; PBS, phosphate-buffered saline.

destroyed the proteolytic activity of 3C^{pro}. In addition, monospecific peptide antisera raised against some of the regions in 3C^{pro} corresponding to the trypsin catalytic triad and specificity pocket, efficiently immunoprecipitated 3C^{pro}. Our results suggest that the picornaviral 3C cysteine proteases and cellular serine proteases may catalyze peptide bond cleavage utilizing basically similar mechanisms.

MATERIALS AND METHODS

Oligonucleotides and Peptides—Oligonucleotides 1 to 3 (Table I) and the sequencing primer 5' GCGTGTGACTGGATTT 3' (HRV-14 nucleotides 5823–5839; Stanway *et al.*, 1984) were synthesized using a Pharmacia Gene Assembler. Oligonucleotides 4 to 9 (Table I) were purchased from Promega. Peptide 1 (CGGGTLDRNEKFRDIR, Fig. 1) and peptide 2 (RYDYATKTGQC, Fig. 1) were purchased from Diagnostic Biotechnology (Singapore) and Cambridge Research Biochemicals (United Kingdom), respectively.

Preparation and Characterization of Peptide Antisera—A non-natural cysteine and three glycine spacers were added to the amino terminus of the core peptide 1 sequence (TLDRNEKFRDIR) to facilitate coupling of the peptide to the carrier protein keyhole limpet hemocyanin (KLH) (Sigma). No additional amino acids were introduced into peptide 2 (RYDYATKTGQC) which already has a cysteine at the carboxyl end. 2.5 mg each synthetic peptide was coupled to KLH via cysteine using *N*-maleimidobenzyl-*N*-hydroxysuccinimide ester (Pierce Chemical Co.) (Nivison and Hanson, 1987).

To induce anti-peptide antibodies, two rabbits were subcutaneously inoculated with 100 µg of each of the KLH-coupled peptides mixed with an equal volume of Freund's complete adjuvant. Subsequent injections were carried out with the same amount of coupled peptides emulsified in Freund's incomplete adjuvant at monthly intervals. Sera were prepared from blood collected 2 weeks after each booster and kept at -70 °C.

For dot blot analysis, serially diluted peptides and KLH were spotted onto nitrocellulose membranes (0.45 µm, Sartorius) and dried. The membranes were incubated with 5% skim milk in phosphate-buffered saline containing 0.05% Tween 20 (PBS-T) at 22 °C for 2 h. The blocked membranes were then incubated with the test sera diluted in PBS-T at 22 °C for 16 h. The membranes were washed three times with PBS-T and incubated with biotinylated goat anti-rabbit IgG (Bethesda Research Laboratories) at 22 °C for 1 h, then washed again three times. The membranes were treated with Streptavidin-horseradish peroxidase conjugate (Bethesda Research Laboratories) at 22 °C for 1 h, washed as before, and incubated with 0.33% 4-chloro-naphthol in methanol and 0.018% hydrogen peroxide in PBS.

Maxicell Labeling and Protein Analysis—Polypeptides expressed by plasmids in *E. coli* maxicell strain CSR603 (Sancar *et al.*, 1979) were labeled with [³⁵S]methionine (>1200 Ci/mmol, Amersham Corp.) according to Cheah *et al.* (1988), except that the cell pellet was resuspended in lysis buffer containing 50 mM Tris-HCl, pH 7.5, 30 mM NaCl, and 200 µg/ml lysozyme. Cell lysis was achieved by three rapid freeze-thaw cycles. The lysed cells were centrifuged for 20 min

at 4 °C and the supernatant (soluble fraction) was saved. The pellet (insoluble fraction) was resuspended in lysis buffer. 5 µl of the soluble and resuspended insoluble fractions were mixed with an equal volume of loading buffer (25 mM Tris-HCl, pH 6.8, 3% SDS, 7.5% β-mercaptoethanol, 25% glycerol, and 0.05% bromophenol blue), boiled for 10 min, subjected to SDS-PAGE, and autoradiographed (Cheah *et al.*, 1988).

Immunoprecipitation—25 µl of antiserum, diluted in 300 µl of immunoprecipitation buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, and 2% Triton X-100), were preabsorbed with KLH and unlabeled *E. coli* maxicell extract at 22 °C for 2 h. 20 µl of [³⁵S]methionine-labeled *E. coli* maxicell extract was then added to the preabsorbed antiserum and mixed at 4 °C for 17 h. 100 µl of protein A-Sepharose CL-4B (Pharmacia LKB Biotechnology Inc.) was added, mixed for a further 1 h, and centrifuged. The pellet was washed three times with immunoprecipitation buffer and 10 mM Tris-HCl, pH 7.5, resuspended in 50 µl of loading buffer, boiled for 10 min, and analyzed by SDS-PAGE.

For the analysis of gel-purified polypeptides, [³⁵S]methionine-labeled polypeptides were separated by SDS-PAGE (Cheah *et al.*, 1988). The gel was rinsed with NT buffer (25 mM Tris-HCl, pH 7.4, and 25 mM NaCl), immediately dried, and autoradiographed. The areas of the gel corresponding to the 3C^{pro} precursor and the 20-kDa 3C^{pro} were cut out and soaked in NT buffer at 4 °C for 17 h. The supernatant, containing diffused proteins, was immunoprecipitated as described above and analyzed by SDS-PAGE.

Site-directed Mutagenesis and DNA Sequencing—The mutagenesis protocol was essentially as described by Kunkel *et al.* (1987) using the Muta-gene[®] M13 *in vitro* mutagenesis kit (Bio-Rad). First a M13 recombinant was constructed, consisting of the entire plasmid pKCC110 (Cheah *et al.*, 1988) subcloned in the *Pst*I site of bacteriophage M13 mp19 to give pLC177. To prevent deletion of the insert, a plaque picked directly from the transformation was grown for 6 h in 6 ml 2 × TY medium, and the single-stranded DNA purified as follows: 5 ml culture supernatant from a 10-min centrifugation was mixed with 0.65 ml of 20% polyethylene glycol 6000 and 2.5 M NaCl. After 15 min at 22 °C, the phage was collected by centrifugation (10 min) and the pellet dissolved in 250 µl of 20 mM Tris-HCl, pH 8.0, 1 mM EDTA. DNA was isolated by two phenol extractions and one chloroform extraction, then precipitated with ethanol.

The template DNA for mutagenesis, uracil-enriched pLC177 single-stranded DNA, was obtained by retransforming the recombinant single-stranded phage DNA (pLC177) into the *Dut*⁻ *Ung*⁻ *E. coli* strain CJ236 (Kunkel *et al.*, 1987), and purifying the single-stranded DNA as above.

The annealing of the mismatching oligonucleotides (Table I) to the template DNA and polymerization with T4 DNA polymerase in the presence of T4 gene 32 protein were performed essentially according to the manufacturer's instructions (Bio-Rad Muta-gene[®] kit), except that the polymerization reaction was incubated at 25 °C for 18 h following the recommended incubations at 4, 25, and 37 °C. The resultant closed, circular DNA was transformed into the *Ung*⁺ *E. coli* strain MV1190 and four independent plaques from each mutagenesis mixture were screened for the correct mutation by dideoxy sequencing

TABLE I
Mutations generated by site-directed mutagenesis

Sequence of mutagenic oligonucleotide 5'→3'	Location of oligonucleotides on HRV-14 cDNA ^a	Amino acid substitution ^b	Predicted role of amino acid ^c
1. CACCTCCAGACTGCCAG	5663–5680	Cys-146→Ser (pAC304)	Catalysis
2. CACAGCACACCTCCCATCTGCCAGTTTTTG	5657–5687	Cys-146→Met (pAC305)	Catalysis
3. CACAGCACACCTCCAGTCTGCCAGTTTTTG	5657–5687	Cys-146→Thr (pAC306)	Catalysis
4. GCTGTGCGTCTGTGGGTATC	5343–5362	His-40→Asp (pAC307)	Catalysis
5. CCCTGATAGCTCTGAATTTTTTC	5476–5497	Asp-85→Ala (pAC308)	Catalysis
6. CCCAGTTTTTGATGCATAATCATAAC	5642–5667	Thr-141→Ser (pAC309)	Base of specificity pocket
7. CAACATGAATATCAAAGATCTTAC	5696–5719	Gly-158→Asp (pAC310)	Highly conserved
8. CGCCAACATTAATACCAAAGATC	5700–5722	His-160→Asn (pAC311)	Side of specificity pocket
9. CTTCATTACCGTCAACATGAATAC	5708–5732	Gly-162→Asp (pAC312)	Top of specificity pocket

^a Nucleotide number shown is based on the published HRV-14 sequence (Stanway *et al.*, 1984).

^b Plasmid names are shown in parenthesis (see text for details).

^c According to the alignment with trypsin (Bazan and Fletterick, 1988).

(Sanger *et al.*, 1977) using the primer 5' GCGTGTGACTGGATT 3'.

To regenerate plasmids equivalent to the parental plasmid pKCC110, the mutant derivatives of pLC177 were digested with *Pst*I (Amersham Corp.), and the linear DNA was allowed to self-ligate. The DNA was transformed into *E. coli* strain MC1022 and ampicillin-resistant (*Ap^R*) transformants were selected (Maniatis *et al.*, 1982). Finally, the mutant plasmid DNAs were retransformed in *E. coli* CSR603 maxicells for analysis of plasmid-encoded proteins (see above).

RESULTS

Immunoprecipitation of 3C^{pro} and Its ~55-kDa Precursor—

The predicted HRV-14 3C^{pro} amino acid sequence (Stanway *et al.*, 1984) was analyzed for short peptide regions with a good potential for inducing antibodies that would recognize surface epitopes in 3C^{pro} (Garnier *et al.*, 1978; Lerner, 1984). The analysis predicted that amino acids 76 to 87 and 136 to 146 (peptides 1 and 2, respectively, Fig. 1) lie in hydrophilic turn regions in the protein, which is in agreement with Werner *et al.* (1986). These peptides were therefore chosen for raising antisera. Two rabbits were independently immunized with each peptide coupled with KLH. Sera from each pair of rabbits reacted with the homologous peptide in a dot blot assay, and no cross-reactivity was detected with the heterologous peptides. Preimmune sera from all four rabbits gave no reaction with either peptide (not shown).

We have previously reported the construction of a HRV-14 expression plasmid pKCC110 which codes for 3C^{pro} plus some flanking viral sequences. In *E. coli* maxicells, pKCC110 encodes a unique precursor polypeptide of ~55-kDa, which was suggested on the basis of its size to comprise the carboxyl-terminal portion of the viral RNA-linked protein VPg (3B), the entire 3C^{pro} and the amino-terminal half of the viral polymerase 3D (3D) (Fig. 1; Cheah *et al.*, 1988). The ~55-kDa 3C^{pro} precursor is rapidly processed to several polypeptides, including 3C^{pro} migrating at ~20 kDa (Cheah *et al.*, 1988).

Fig. 2A shows that in extracts of [³⁵S]methionine-labeled *E. coli* maxicells harboring pKCC110, 3C^{pro} and the ~55-kDa 3C^{pro} precursor are more abundant in the insoluble pellet than

in the lysozyme (soluble) extract (Fig. 2A, compares lanes 2 and 3). A background protein comigrating with the ~55-kDa band is occasionally detected in the soluble fraction of maxicells carrying the vector pKCC100 (Fig. 2A, lane 4).

Immunoprecipitation experiments using the soluble fraction (lysozyme supernatant; Fig. 2A, lane 3) demonstrated that peptide 1 and 2 antisera specifically recognize the 20-kDa 3C^{pro} polypeptide (Fig. 2B, lanes 2 and 5), whereas the preimmune sera did not (Fig. 2B, lanes 3 and 6). The ~55-kDa 3C^{pro} precursor from the soluble fraction of *E. coli* was not immunoprecipitated by either peptide antisera (Fig. 2B, lanes 2 and 5).

To circumvent the lack of immunoprecipitation of the ~55-kDa 3C^{pro} precursor protein, the [³⁵S]methionine-labeled proteins encoded by pKCC110 in *E. coli* maxicells were separated by SDS-PAGE, and the gel was immediately dried and autoradiographed without fixing the proteins. The regions corresponding to the ~55-kDa 3C^{pro} precursor and 3C^{pro} (Fig. 2A, lane 1) were excised from the dried gel and eluted by diffusion at 4 °C. The eluted proteins were either rerun on a second SDS-polyacrylamide gel (Fig. 2C, lanes 1 and 6) or incubated with peptide 1 and 2 antisera and immunoprecipitated. Both peptide antisera immunoprecipitated the ~55-kDa 3C^{pro} precursor (Fig. 2C, lanes 2 and 3) and 3C^{pro} (Fig. 2C, lanes 7 and 8), whereas preimmune sera did not (Fig. 2C, lanes 4, 5, 9, and 10). Further, the immunoprecipitation of the gel-purified 3C^{pro} precursor by both peptide antisera was inhibited by prior absorption of the peptide antisera with 10 µg of the homologous peptide (not shown).

Taken together, the immunoprecipitation experiments confirmed our previous assignment of the ~55- and ~20-kDa polypeptides as 3C^{pro} precursor and 3C^{pro}, respectively (Cheah *et al.*, 1988) and clearly indicate that amino acids 76 to 87 and 136 to 146 are surface epitopes of 3C^{pro} (Fig. 1).

Construction of 3C^{pro} Mutants—Computer alignments of animal picornavirus 3C proteases and cellular serine proteases have indicated a limited number of significant homologies. The presumed active-site Cys-147 of poliovirus 3C^{pro}, equiv-

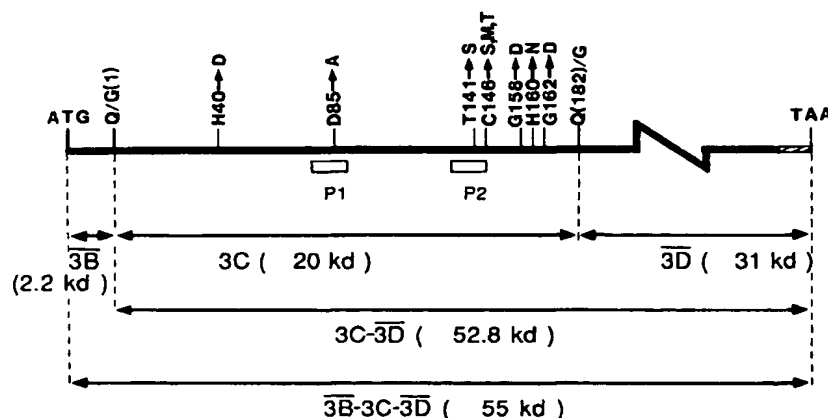


FIG. 1. Schematic diagram showing the HRV-14 portion of recombinant plasmid pKCC110. The heavy blackened line represents the cDNA of HRV-14 cloned in the *trp* promoter expression vector pKCC100, and the hatched box depicts the 19 amino acids derived from vector sequences fused in frame to the HRV-14 open reading frame (Cheah *et al.*, 1988). The proposed Gln/Gly cleavage sites flanking 3C^{pro} are shown as Q/G(1) and Q(182)/G (Stanway *et al.*, 1984; Cheah *et al.*, 1988). Peptide sequences chosen for raising antibodies, shown as open boxes, are P1 (peptide 1, amino acids 76 to 87 with an amino-terminal extension of Cys-Gly-Gly-Gly) and P2 (peptide 2, amino acids 136 to 146). The full sequences of the peptides are given under "Materials and Methods." The locations of the amino acids substituted by site-directed mutagenesis are shown in single letter code (see text and Table I for details). The viral proteins and their precursors (3B, 3C, 3D, 3C-3D, and 3B-3C-3D) are shown with the estimated sizes in parentheses (Stanway *et al.*, 1984; Cheah *et al.*, 1988). Truncated proteins are indicated by overlining (e.g. 3D).

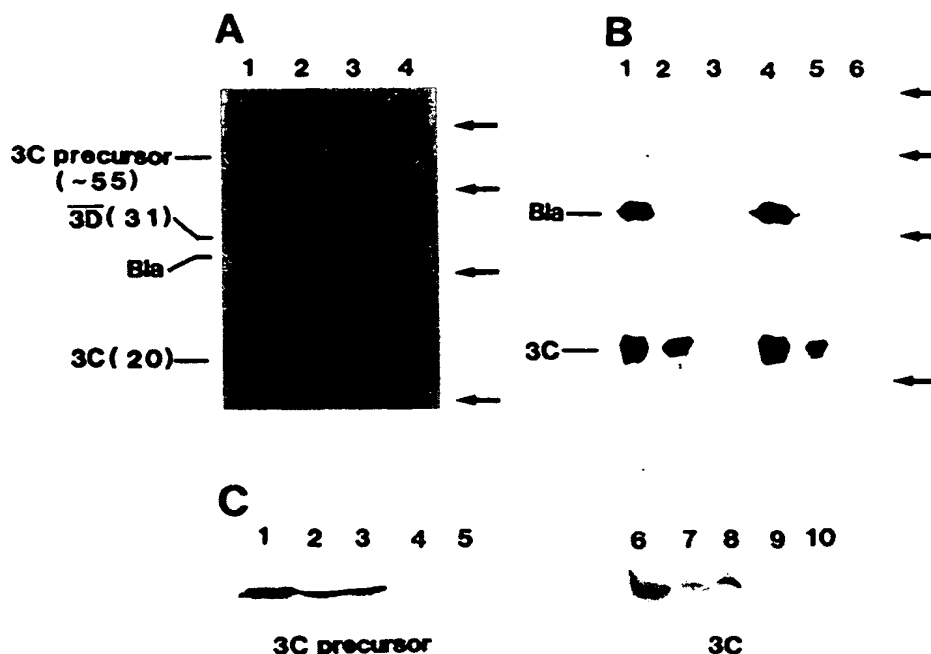


FIG. 2. Protein analysis. A, autoradiograph of a 12.5% SDS-polyacrylamide gel showing [35 S]methionine-labeled HRV-14 polypeptides synthesized in *E. coli* CSR603 maxicells. Lane 1, pKCC110 (whole lysate); lane 2, pKCC110 (solubilized pellet fraction); lane 3, pKCC110 (soluble fraction extracted with lysozyme); lane 4, vector pKCC100 without insert (soluble fraction extracted with lysozyme). Unique polypeptides encoded by recombinant plasmid pKCC110 are indicated on the left (Fig. 1; Cheah *et al.*, 1988). Bla is β -lactamase. B, immunoprecipitation of protease 3C by peptide antisera. [35 S]Methionine-labeled soluble proteins encoded by pKCC110 were either loaded directly on the SDS-polyacrylamide gel (lanes 1 and 4), immunoprecipitated with peptide 1 antiserum (lane 2), or immunoprecipitated with peptide 2 antiserum (lane 5). Lanes 3 and 6 are identical to lanes 2 and 5, respectively, except that preimmune sera were used. The arrowheads on the right of panels A and B indicate the positions of size standards from top to bottom of sizes 68, 43, 25.7, and 18.4 kDa. C, immunoprecipitation of SDS-polyacrylamide gel-purified 3C^{pro} precursor (left panel) and 3C^{pro} (right panel). The regions in the gel (Fig. 2A, lane 1) corresponding to the 3C^{pro} precursor and 3C^{pro} were excised, and the proteins were eluted and analyzed on a 12.5% SDS-polyacrylamide gel. Lanes 1 and 6, proteins loaded directly; lanes 2 and 7, immunoprecipitation with peptide 1 antiserum; lanes 3 and 8, immunoprecipitation with peptide 2 antiserum; lanes 4, 5, 9, and 10, immunoprecipitation with preimmune sera.

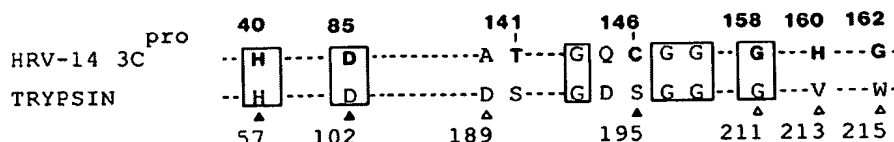


FIG. 3. Proposed alignment of catalytic and specificity pocket amino acids of trypsin and HRV-14 3C^{pro}. Computer alignment of the catalytic triad (Δ) and specificity pocket (Δ) amino acids of trypsin with the corresponding residues of HRV-14 3C^{pro} is shown (Bazan and Fletterick, 1988). Amino acids in HRV-14 3C^{pro} substituted by site-directed mutagenesis (Fig. 1, Table I) are shown in *bold type*. Based on our results, T-141 and not A-140 of 3C^{pro} may be equivalent to D-189 of trypsin (see "Discussion"). Identical amino acids are boxed.

alent to Cys-146 in HRV-14 3C^{pro}, is highly conserved in all animal picornaviruses and lies in an area of significant homology with the active-site Ser-195 of trypsin-like serine proteases (Gorbalenya *et al.*, 1986; Bazan and Fletterick, 1988). In addition, His-40 and Asp-85 of HRV-14 (Stanway *et al.*, 1984) are highly conserved in animal picornaviruses and cellular serine proteases. His-40, Asp-85, and Cys-146 of HRV-14 can be superimposed on the trypsin serine protease catalytic triad, His-57, Asp-102, and Ser-195 (Fig. 3; Kraut, 1977; Craik *et al.*, 1987; Sprang *et al.*, 1987). Therefore, substitutions were made individually at His-40 and Asp-85, and three different substitutions were made at Cys-146 to test whether these amino acids are essential for the catalytic function of 3C^{pro} (Table I, Fig. 1).

The computer alignments also revealed that HRV-14 3C^{pro} amino acids Thr-141, His-160, and Gly-162 lie in positions equivalent to serine protease amino acids known to be important for substrate binding and specificity (Fig. 3; Kraut, 1977; Bazan and Fletterick, 1988). In trypsin, the equivalent amino acids are serine, valine, and tryptophan, respectively (Fig. 3). Thr-141 and His-160 are highly conserved in picornaviruses, while Gly-162 is only partially conserved. Two lines of evidence suggest that these 3 residues are among those which are important determinants of Gln-Gly cleavage specificity. First, molecular modeling of His-160/Gly-162 in the pocket of a trypsin-inhibitor complex structure revealed possible hydrogen-bonding interactions between viral Thr-141/His-160 and the enzyme-bound side chain of the Gln substrate



FIG. 5. Polypeptides encoded by protease 3C mutant plasmids. The [35 S]methionine-labeled polypeptides in the whole extracts of *E. coli* CSR603 harboring various recombinant plasmids (Table 1) were separated by SDS-PAGE. Lane 1, the vector pKCC100; lane 2, pKCC110 (parent); lane 3, pAC304 (Cys-146 to Ser); lane 4, pAC305 (Cys-146 to Met); lane 5, pAC306 (Cys-146 to Thr); lane 6, pKCC110 (parent); lane 7, pAC307 (His-40 to Asp); lane 8, pAC310 (Gly-158 to Asp); lane 9, pAC311 (His-160 to Asn); lane 10, pAC312 (Gly-162 to Asp); lane 11, pAC308 (Asp-85 to Ala); lane 12, pAC309 (Thr-141 to Ser). Arrows on the right show the positions of protein markers with sizes from top to bottom of 68, 43, 25.7, and 18.4 kDa. Indicated on the left are the pKCC110-encoded viral polypeptides, 3B-3C-3D (55 kDa), 3C-3D (52.8 kDa), 3D (31 kDa), and 3C (20 kDa) (Fig. 1). Bla is β -lactamase.

4-h chase with unlabeled methionine and chloramphenicol, nearly all the parental 3C^{pro} precursor was processed to 3D and 3C^{pro} (see also Fig. 5 of Cheah *et al.*, 1988). In contrast, no processing of the 3B-3C-3D precursor to 3C-3D, 3D, and 3C^{pro} was detected with the Asp-85 to Ala mutant, even during an 18-h chase period (Fig. 6B). An identical result was obtained with the His-40 to Asp, Cys-146 to Ser, Cys-146 to Met, Cys-146 to Thr, Gly-158 to Asp, His-160 to Asn, and Gly-162 to Asp mutants (not shown). With the Thr-141 to Ser mutant, the 3B-3C-3D/3C-3D doublet was processed during the chase period to 3D and a 3C^{pro} mutant polypeptide (Fig. 6C), albeit at a much slower rate than that of the parental 3C^{pro} precursor (Fig. 6A). These results strengthen our conclusion that mutations at six amino acid positions totally inactivate 3C^{pro}, and mutation of Thr-141 to Ser severely impairs 3C proteolytic activity.

DISCUSSION

We have previously utilized the *E. coli* maxicell system to demonstrate expression and autocatalytic proteolysis of an HRV-14 3C^{pro} precursor (Cheah *et al.*, 1988). In the present study, the parental and mutant 3C^{pro} precursors were expressed at comparable levels in *E. coli* maxicells, but the parental precursor migrated slightly faster in denaturing gels than the proteolytically inactive mutant precursors (Fig. 5). This is because cleavage of the parental 3B-3C-3D precursor is much faster at the 3B/3C junction than at the 3C/3D junction, resulting in the accumulation of a 3C-3D precursor of 52.8 kDa (Fig. 1). In other picornaviruses, cleavage at 3B/3C has also been reported to be faster than cleavage at 3C/3D (Strebel *et al.*, 1986; Richards *et al.*, 1987; Jore *et al.*, 1988). *In vivo*, a slow cleavage at 3C/3D would control the release of mature 3C^{pro} and at the same time provide an adequate supply of 3C-3D, the active protease required for cleavage of the capsid protein precursors (Jore *et al.*, 1988; Ypma-Wong *et al.*, 1988).

The *E. coli* maxicell system has for the first time provided a sensitive, convenient, and rapid way of assaying the effects of single amino acid substitutions on the proteolytic activity

of autocatalytic proteases. Seven amino acid positions in HRV-14 3C^{pro} were chosen for site-directed mutagenesis based on two considerations. First, amino acids at all seven positions are highly conserved in animal picornaviruses. Second, an alignment with trypsin predicted that certain 3C^{pro} residues may be involved either in catalysis or substrate binding and specificity (Fig. 3; Bazan and Fletterick, 1988). It has previously been shown that the Cys-147 to Ser mutation inactivates poliovirus 3C^{pro}, although it was not clear whether residual proteolytic activity remained (Ivanoff *et al.*, 1986). Here we show that if Cys-146 of HRV-14 3C^{pro} (equivalent to poliovirus Cys-147) was changed either to serine, methionine, or threonine, proteolytic activity was completely destroyed. Likewise, mutation of His-40 to Asp or Asp-85 to Ala, which are equivalent to His-57 and Asp-102 in the catalytic triad of the trypsin-like serine proteases, completely destroyed 3C^{pro} activity. Two different antisera raised against peptides containing 3C^{pro} amino acids 76 to 87 and 136 to 146 efficiently immunoprecipitated mature 3C^{pro}, strongly suggesting that Asp-85 and Cys-146 lie in accessible surface locations in 3C^{pro}. Taken together, the site-directed mutagenesis and immunoprecipitation data suggest that catalysis by HRV-14 3C^{pro} is performed by a surface triad of His-40, Asp-85, and Cys-146 in a mechanistically similar fashion to the histidine, aspartic acid, and serine at the active-site of the trypsin-like serine proteases (Fig. 3; Kraut, 1977; Craik *et al.*, 1987).

A very recent independent alignment of viral cysteine and cellular serine proteases (Gorbalenya *et al.*, 1989) is largely in agreement with the analysis of Bazan and Fletterick (1988), except that Glu-71 and not Asp-85 was suggested to represent the acidic amino acid in the catalytic triad of HRV-14 and most other picornavirus 3C proteases. Although a glutamic acid has never been found in the serine protease catalytic triad and some 3C proteases have Asp-71, the participation of position 71 in the catalytic triad of 3C cysteine proteases cannot be ruled out.

Amino acids in viral 3C proteases predicted to be involved in determining Gln-Gly cleavage specificity include the HRV-14 residues Ala-140, Thr-141, Gly-158, His-160, and Gly-162

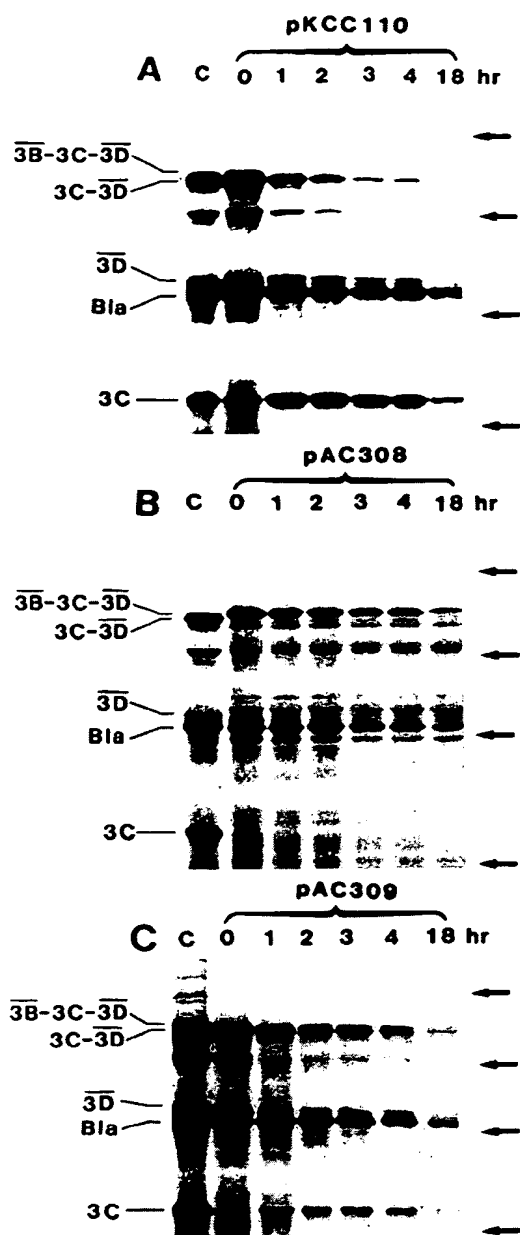


FIG. 6. Kinetics of cleavage of parent and protease 3C precursors. Viral polypeptides expressed in UV-irradiated *E. coli* maxicells were labeled with [35 S]methionine for 2 min and chased for the times indicated at 37 °C in the presence of excess unlabeled methionine and chloramphenicol (Cheah *et al.*, 1988). Panel A, pKCC110 (parent); panel B, pAC308 (Asp-85 to Ala); panel C, pAC309 (Thr-141 to Ser). Arrows show the positions of protein markers with sizes from top to bottom of 68, 43, 25.7, and 18.4 kDa. Indicated on the left are the viral polypeptides, 3B-3C-3D (55 kDa), 3C-3D (52.8 kDa), 3D (31 kDa), and 3C (20 kDa) (Fig. 1). Bla is β -lactamase.

(Fig. 3; Bazan and Fletterick, 1988; Gorbalenya *et al.*, 1989). Ala-140 in HRV-14 3C^{pro} aligns with Asp-189 of trypsin, an important determinant of Arg/Lys cleavage specificity located at the base of the substrate binding pocket (Graf *et al.*, 1987). However, Ala-140 is unlikely to be directly involved in 3C^{pro} specificity, since other picornaviruses have the functionally dissimilar residues Gln, Asn, Glu, or Pro in this position. We

found that Gly-158 to Asp, His-160 to Asn, and Gly-162 to Asp substitutions abolished 3C^{pro} activity, supporting the proposal that each of the amino acids in these positions plays a crucial role in cleavage specificity (Bazan and Fletterick, 1988). Consistent with our results, the His-161 of poliovirus 3C^{pro} (equivalent to His-160 of HRV-14) was converted to a glycine and proteolytic activity was also lost (Ivanoff *et al.*, 1986). The Thr-141 to Ser mutation in HRV-14 3C^{pro} markedly reduced its activity. Our immunoprecipitation data suggest that Thr-141 lies in an accessible surface region and, as discussed earlier, Thr-141 could form a hydrogen bond with the side chain of the S1-bound Gln substrate. In theory, Ser-141 could similarly form a hydrogen bond, but the interaction would be weaker, since serine has a shorter side chain than threonine. A weaker interaction might explain the impaired activity of the Ser-141 mutant. Based on these considerations, we speculate that Thr-141 and not Ala-140 of 3C^{pro} is equivalent to the important Asp-189 of trypsin (Fig. 3; Graf *et al.*, 1987).

It is remarkable that substitutions at six positions in 3C^{pro} completely destroyed proteolytic activity, and one additional substitution (Thr-141 to Ser) severely impaired activity. It could be argued that 3C proteases are highly sensitive to structural changes. Although we cannot exclude this possibility, there are two considerations which argue against it. First, some substitutions in poliovirus 3C^{pro} are without effect (Ivanoff *et al.*, 1986; Dewalt and Semler, 1987). Second, the 3C proteases of two related HRV subtypes HRV-2 and HRV-14 are less than 50% homologous, and structurally dissimilar amino acids align at many positions (Stanway *et al.*, 1984; Skern *et al.*, 1985).

We have demonstrated that seven amino acids which are highly conserved in the 3C proteases of animal picornaviruses are important for the proteolytic activity of HRV-14 3C^{pro}. These amino acids align with catalytic or specificity pocket residues of trypsin, suggesting that the catalytic mechanism utilized by picornavirus 3C cysteine proteases is closely related to that of the cellular trypsin-like serine proteases. This is interesting because trypsin and chymotrypsin are inactive as precursors, which is in sharp contrast to the viral 3C proteases. Also, unlike the cellular serine proteases, the viral 3C cysteine proteases are believed to cleave both in *cis* and in *trans* (Kräusslich and Wimmer, 1988). The question of whether the mechanisms of *cis* and *trans* catalysis are different has not yet been addressed.

If the 3C cysteine proteases and cellular serine proteases are structurally and functionally related, it may be possible to convert a viral 3C cysteine protease to a serine protease by substituting a limited set of amino acids to compensate for the Cys-146 to Ser change, which by itself inactivates 3C^{pro}. Support for this concept comes from the observation mentioned earlier that *S. aureus* (strain V8) protease is a serine protease which cleaves after Glu residues and has a Thr-141/His-160/Gly-162 complement of amino acids in the substrate-binding pocket (Drapeau, 1978; Bazan and Fletterick, 1988). In addition, animal flaviviruses and pestiviruses code for 3C^{pro}-like serine proteases with Arg/Lys cleavage specificity and only limited homology with the trypsin class of serine proteases in and around the substrate-binding pocket (Bazan and Fletterick, 1989).

In conclusion, our site-directed mutagenesis results combined with a knowledge of the physicochemical properties of purified 3C proteases together with x-ray crystal structure data, will lead to a better understanding of the catalytic mechanism utilized by this unusual class of proteases.

Acknowledgments—We are grateful to Dr. Gerd Klock and Woon-

Khiong Chan for critically reading the manuscript, Sabita Sankar for assistance in the preparation of peptide antisera, Mei-Yeng Kok for oligonucleotide synthesis, Ka-Liong Lok for photography, and Azizah Mohd Ali for typing the manuscript.

REFERENCES

- Argos, P., Kamer, G., Nicklin, M. J. H., and Wimmer, E. (1984) *Nucleic Acids Res.* **12**, 7251-7267
- Bazan, J. F., and Fletterick, R. J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7872-7876
- Bazan, J. F., and Fletterick, R. J. (1989) *Virology* **171**, 637-639
- Cheah, K.-C., Sankar, S., and Porter, A. G. (1988) *Gene (Amst.)* **69**, 265-274
- Craik, C. S., Rocznick, S., Largman, C., and Rutter, W. J. (1987) *Science* **237**, 909-913
- Dewalt, P. G., and Semler, B. L. (1987) *J. Virol.* **61**, 2162-2170
- Drapeau, G. R. (1978) *J. Bacteriol.* **136**, 607-613
- Franssen, H., Leunissen, J., Goldbach, R., Lomonosoff, G., and Zimmermann, D. (1984) *EMBO J.* **3**, 855-861
- Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) *J. Mol. Biol.* **120**, 97-120
- Gorbalenya, A. E., Blinov, V. M., and Donchenko, A. P. (1986) *FEBS Lett.* **194**, 253-257
- Gorbalenya, A. E., Donchenko, A. P., Blinov, V. M., and Koonin, E. V. (1989) *FEBS Lett.* **243**, 103-114
- Graf, L., Craik, C. S., Patthy, A., Rocznick, S., Fletterick, R. J., and Rutter, W. J. (1987) *Biochemistry* **26**, 2616-2623
- Gwaltney, J. M. (1975) *Yale J. Biol. Med.* **48**, 17-45
- Hanecak, R., Semler, B. L., Ariga, H., Anderson, C. W., and Wimmer, E. (1984) *Cell* **37**, 1063-1073
- Ivanoff, L. A., Towatari, T., Ray, J., Korant, B. D., and Petteway, S. R. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83**, 5392-5396
- Jore, J., De Geus, B., Jackson, R. J., Pouwels, P. H., and Enger-Valk, B. E. (1988) *J. Gen. Virol.* **69**, 1627-1636
- Klump, W., Marquardt, O., and Hofschneider, P. H. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 3351-3355
- Korant, B. (1973) *J. Virol.* **12**, 556-563
- Korant, B. D., Brzin, J., and Turk, V. (1985) *Biochem. Biophys. Res. Commun.* **127**, 1072-1076
- Kräusslich, H.-G., and Wimmer, E. (1988) *Annu. Rev. Biochem.* **57**, 701-754
- Kraut, J. (1977) *Annu. Rev. Biochem.* **46**, 331-358
- Kunkel, T. A., Roberts, J. D., and Zabour, R. A. (1987) *Methods Enzymol.* **154**, 367-382
- Lerner, R. A. (1984) *Adv. Immunol.* **36**, 1-44
- Libby, R. T., Cosman, D., Cooney, M. K., Merriam, J. E., March, C. J., and Hopp, T. P. (1988) *Biochemistry* **27**, 6262-6268
- Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
- Nicklin, M. J. H., Toyoda, H., Murray, M. G., and Wimmer, E. (1986) *Bio/Technology* **4**, 33-42
- Nicklin, M. J. H., Harris, K. S., Pallai, P. V., and Wimmer, E. (1988) *J. Virol.* **62**, 4586-4593
- Nivison, H. T., and Hanson, M. R. (1987) *Plant Mol. Biol. Rep.* **5**, 295-309
- Parks, G. D., Baker, J. C., and Palmenberg, A. C. (1989) *J. Virol.* **63**, 1054-1058
- Pelham, H. R. B. (1978) *Eur. J. Biochem.* **85**, 457-462
- Richards, O. C., Ivanoff, L. A., Bienkowska-Szewczyk, K., Butt, B., Petteway, S. R., Rothstein, M. A., and Ehrenfeld, E. (1987) *Virology* **161**, 348-356
- Sancar, A., Hack, A. M., and Rupp, W. D. (1979) *J. Bacteriol.* **137**, 692-693
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467
- Skern, T., Sommergruber, W., Blaas, D., Gruendler, P., Fraundorfer, F., Pieler, C., Fogy, I., and Kuechler, E. (1985) *Nucleic Acids Res.* **13**, 2111-2126
- Sprang, S., Standing, T., Fletterick, R. J., Stround, R. M., Finer-Moore, J., Xuong, N.-H., Hamlin, R., Rutter, W. J., and Craik, C. S. (1987) *Science* **237**, 905-908
- Stanway, G., Hughes, P. J., Mountford, R. C., Minor, P. D., and Almond, J. W. (1984) *Nucleic Acids Res.* **12**, 7859-7875
- Stott, E. J., and Killington, R. A. (1972) *Annu. Rev. Microbiol.* **26**, 503-524
- Strebel, K., Beck, E., Strohmaier, K., and Schaller, H. (1986) *J. Virol.* **57**, 983-991
- Werner, G., Rosenwirth, B., Bauer, E., Seifert, J.-M., Werner, F.-J., and Besemer, J. (1986) *J. Virol.* **57**, 1084-1093
- Ypma-Wong, M. F., Dewalt, P. G., Johnson, V. H., Lamb, J. G., and Semler, B. L. (1988) *Virology* **166**, 265-270

Exhibit 10



The Catalytic Role of the Active Site Aspartic Acid in Serine Proteases

Charles S. Craik; Steven Rocznik; Corey Largman; William J. Rutter

Science, New Series, Vol. 237, No. 4817 (Aug. 21, 1987), 909-913.

Stable URL:

<http://links.jstor.org/sici?sici=0036-8075%2819870821%293%3A237%3A4817%3C909%3ATCROTA%3E2.0.CO%3B2-5>

Science is currently published by American Association for the Advancement of Science.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/aaas.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

<http://www.jstor.org/>
Thu Sep 9 18:08:55 2004

squares with the computer program CORELS (10). The positional parameters of individual atoms were then refined subject to stereochemical restraints by using the subcell data (6). The positions of missing side-chain atoms and those of the benzimidazole and calcium were determined from the subcell difference electron density map computed from the refined model. A model of the full crystallographic asymmetric unit in the correct $P2_12_12_1$ unit cell was then constructed by adding a replicate of the trypsin molecule translated by 46 Å along the b and 32 Å along c . The full model was refined in three stages. In each stage the model was refit to a difference Fourier map computed with the coefficients ($2F_{\text{obs}} - F_{\text{calc}}$). Strong peaks in the electron density in positions consistent with hydrogen bond contacts to the protein or other established solvent positions were included in the model as ordered solvent. Next, the positional and thermal parameters of all atoms were refined by iterations of restrained crystallographic least squares, with data in the resolution range $6 \text{ Å} \leq d \leq 2.3 \text{ Å}$. Refinement was stopped when further cycles failed to reduce the crystallographic R factor and when the mean shift in coordinate positions was less than 0.05 Å. Refined coordinates were then used to compute phases for a new electron map to be used in the next stage of manual refitting. After the third stage (R factor = 0.18), examination of the electron density failed to reveal errors or ambiguity in main- or side-chain positions, although the side chains of six residues located at the surface of the molecules were disordered and could not be defined. Up to this point, side-chain atoms for His³⁷, Asn¹⁰², or Ser¹⁹⁵ had been excluded from the model. A difference electron density map ($F_{\text{obs}} - F_{\text{calc}}$) revealed strong and well-ordered density for the Asn¹⁰² and Ser¹⁹⁵, but the His³⁷ residue appeared to be statistically disordered (Fig. 2, top) (11).

10. J. L. Sussman, S. R. Holbrook, G. M. Church, S. H. Kim, *Acta Crystallogr.* A32, 311 (1976).
11. The possibility that one or other of the peaks are artifactual was tested by independent refinement of two alternative models: one with His³⁷ fit to the stronger, internal density and the second with His³⁷ fit to the external density. In each model the His³⁷ atoms were assigned full occupancy and side-chain positions for Asn¹⁰² and Ser¹⁹⁵ were included. Each model was subjected to restrained crystallographic refinement by varying the thermal and positional parameters of all atoms. Subsequently, a difference Fourier map ($F_{\text{obs}} - F_{\text{calc}}$) was computed for each model with the use of the refined positional and thermal parameters for all of the atoms in the respective models. In both cases, residual electron density appeared at the alternative histidine site. Again, the observed density peaks were contiguous with the C β atom of His³⁷ and thus could not be interpreted as ordered water molecules. The relative occupancy of the two histidine positions and the total occupancy of both positions relative to other histidine side chains was estimated by integration of difference electron density at all of the histidine side-chain positions in one of the trypsin molecules in the asymmetric unit. The difference Fourier map ($F_{\text{obs}} - F_{\text{calc}}$) used in the integration was computed from a model in which the side-chain atoms of all four histidine residues (at sequence positions 40, 57, 70, and 87) were removed from the coordinate set of one molecule. Integration was performed manually by summing over all grid points within 2.0 Å of histidine atomic positions that had electron density at least one standard deviation greater than the background density. After normalization the apparent relative integrated difference densities at the histidine side-chain positions were: His⁴⁰, 0.87; His⁵⁷, 0.60; His⁷⁰, 0.79; and His⁸⁷, 1.0. All but His³⁷ are well ordered, so the range in integrated densities reflects thermal motion and experimental error. The sum of the density over the two His³⁷ side-chain sites is lower than the mean density of the well-ordered histidine side chains, but is consistent with the high B factors of His³⁷ atoms at both positions. The relative occupancy of the alternative His³⁷ positions was estimated by integrating the difference density at the N δ 1 and C ϵ 1 atoms of the gauche conformer and the C δ 2 and N ϵ 2 atoms of the trans conformer and by taking the ratio of the

integrated densities for the two positions. The remaining histidine atoms were not included in the integration because the resolution of the data set did not allow the densities of the two conformers to be resolved at those positions.

- Final refined positional and thermal parameters for both trans and gauche conformers were determined by refining an atomic model in which both conformers were simultaneously included. Side-chain atoms of the gauche conformer were assigned occupancies of 0.67 and atoms of the trans isomer were assigned occupancies of 0.33 based on the estimate derived from the integration described above (12). After three final cycles of refinement of all thermal and positional parameters of both trypsin monomers in the asymmetric unit, the crystallographic R factor was 0.161.
12. A modified version of PROTEIN (obtained from J. Smith) does not generate restraints between alternate side-chain positions of a statistically disordered residue. This allows refinement of two conformations of an amino acid simultaneously.
13. W. Bode and P. Schwager, *J. Mol. Biol.* 98, 693 (1975).
14. R. Henderson, *ibid.* 54, 341 (1970).
15. An upper estimate of the mean error in atomic position is 0.25 Å. It was obtained by an analysis of the variation of crystallographic R factor as a function of resolution (16).
16. V. Luzatti, *Acta Crystallogr.* 6, 142 (1953).
17. A. A. Kossiakoff and S. A. Spencer, *Biochemistry* 20,

- 6462 (1981).
18. M. Krieger *et al.*, *ibid.* 15, 3458 (1976).
19. M. N. G. James, A. R. Sielecki, G. D. Brayer, L. T. Delbaere, C. A. Bauer, *J. Mol. Biol.* 144, 43 (1980).
20. P. H. Morgan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 69, 3312 (1972).
21. A. A. Kossiakoff *et al.*, *Biochemistry* 16, 654 (1977); H. Fehlbauer, W. Bode, R. Huber, *J. Mol. Biol.* 111, 415 (1977).
22. J. L. Chambers *et al.*, *Biochem. Biophys. Res. Commun.* 59, 70 (1974).
23. M. O. Jones and R. M. Stroud, *Biochemistry*, in press.
24. D. M. Blow *et al.*, *Nature (London)* 221, 337 (1969).
25. C. S. Craik *et al.*, *J. Biol. Chem.* 259, 14255 (1984).
26. The coordinates were obtained from the Protein Data Bank at Brookhaven National Laboratory.
27. We thank J. Sadowsky, C. Nielsen, and E. Goldsmith for assistance with Area Detector data collection and processing and B. Montfort for assistance with crystallographic refinement calculations. We gratefully acknowledge grant support from NIH: AM31507 to S.R.S., GM24485 to R.M.S., and AM26081 to R.J.F.; from NSF: DMB8608086 to C.S.C. and PCM830610 to W.J.R.; a Bristol Meyer grant of Research Corporation and a CCRC grant to C.S.C. The coordinates of the D 102 N trypsin structure at pH 6 have been submitted to the Protein Data Bank at Brookhaven National Laboratory.

29 September 1986; accepted 29 May 1987

The Catalytic Role of the Active Site Aspartic Acid in Serine Proteases

CHARLES S. CRAIK, STEVEN ROCZNIAK,* COREY LARGMAN,† WILLIAM J. RUTTER

The role of the aspartic acid residue in the serine protease catalytic triad Asp, His, and Ser has been tested by replacing Asp¹⁰² of trypsin with Asn by site-directed mutagenesis. The naturally occurring and mutant enzymes were produced in a heterologous expression system, purified to homogeneity, and characterized. At neutral pH the mutant enzyme activity with an ester substrate and with the Ser¹⁹⁵-specific reagent diisopropylfluorophosphate is approximately 10⁴ times less than that of the unmodified enzyme. In contrast to the dramatic loss in reactivity of Ser¹⁹⁵, the mutant trypsin reacts with the His⁵⁷-specific reagent, tosyl-L-lysine chloromethylketone, only five times less efficiently than the unmodified enzyme. Thus, the ability of His⁵⁷ to react with this affinity label is not severely compromised. The catalytic activity of the mutant enzyme increases with increasing pH so that at pH 10.2 the k_{cat} is 6 percent that of trypsin. Kinetic analysis of this novel activity suggests this is due in part to participation of either a titratable base or of hydroxide ion in the catalytic mechanism. By demonstrating the importance of the aspartate residue in catalysis, especially at physiological pH, these experiments provide a rationalization for the evolutionary conservation of the catalytic triad.

SERINE PROTEASES FUNCTION IN many biological systems to hydrolyze specific polypeptide bonds. Trypsin, a well-studied member of this family, catalyzes the hydrolysis of peptide and ester substrates that contain lysyl or arginyl side chains. Serine proteases have the triad of residues Asp¹⁰², His⁵⁷, and Ser¹⁹⁵ at the active site (chymotrypsin numbering system). X-ray crystallographic studies reveal that these three residues are in close proximity, which suggests they may serve as a functional interacting unit responsible for bond formation and cleavage during catalysis (1). Numerous chemical and physical

studies indicate that Ser¹⁹⁵ and His⁵⁷ play crucial roles in catalysis. For example, selective reaction of Ser¹⁹⁵ with diisopropylfluor-

C. S. Craik, Departments of Pharmaceutical Chemistry and of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0446.
S. Rocznik, C. Largman, W. J. Rutter, Hormone Research Institute and Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0448.

*Present address: NutraSweet Company, Mount Prospect, IL 60056.

†Present address: Veterans Administration Hospital, Martinez, CA 94553, and Departments of Internal Medicine and Biological Chemistry, University of California, Davis, CA 95616.

ophosphate (DFP) (2) or modification of the His⁵⁷ of trypsin with tosyl-L-lysine chloromethyl ketone (TLCK) (3) blocks catalytic activity. The collective data suggest that substrate hydrolysis is facilitated through nucleophilic attack by the Ser¹⁹⁵ hydroxyl oxygen on the carbonyl carbon of the substrate. Concomitantly the hydroxyl proton of the serine can be transferred to the imidazole of His⁵⁷ and subsequently donated to the resulting leaving group (alcohol or amine) in the reaction. The remaining acyl enzyme intermediate is hydrolyzed by a mechanism that is the reverse of its formation except that water instead of Ser¹⁹⁵ serves as the nucleophile. The role of the buried carboxylate of Asp¹⁰² in the catalytic process remains to be clarified experimentally.

The geometric relation of the amino acids

Table 1. Ratios of activity for trypsin and D 102 N trypsin. Assays for Z-Lys-S-Bzl were performed at pH 7.15 and 10.18 (see legend to Fig. 1 for a description of the experimental conditions). Values for $k_{\text{obs}}/[I]$ with DFP were determined by the method of Kitz and Wilson (24). Standard conditions (25) were used except when the initial DFP concentration was 10 mM in assays with D 102 N trypsin at pH 10.03; background hydrolysis of DFP was relatively rapid and enzymatic activity at infinite times did not equal zero. In this case the $k_{\text{obs}}/[I]$ value (where $[I]$ is the concentration of inhibitor) was determined by the method of Yosogimura *et al.* (26). Values of $k_{\text{obs}}/[I]$ from assays with trypsin were calculated to be $790 \pm 80 \text{ M}^{-1} \text{ min}^{-1}$ (pH 7.96) and $980 \pm 70 \text{ M}^{-1} \text{ min}^{-1}$ (pH 10.03). In assays with D 102 N trypsin these values were $0.070 \pm 0.008 \text{ M}^{-1} \text{ min}^{-1}$ (pH 7.96) and $0.098 \pm 0.019 \text{ M}^{-1} \text{ min}^{-1}$ (pH 10.03). Titrations with MUGB were followed at 360 nm on a Perkin-Elmer LSS spectrofluorometer and performed in triplicate in 50 mM Hepes buffer, pH 7.5, that contained 2 μM MUGB. Titrations of trypsin were complete in 2 seconds (the minimum detection time of the fluorometer) or less when enzyme concentrations ranged from 50 nM to 400 nM. Approximately 17 minutes elapsed before a molar equivalence of MUGB reacted with 400 nM D 102 N trypsin. Values for $k_{\text{obs}}/[I]$ with TLCK were determined by the method of Kitz and Wilson (24); standard conditions were used (27). $k_{\text{obs}}/[I]$ values from assays with trypsin were calculated to be $760 \text{ M}^{-1} \text{ min}^{-1}$ (pH 7.16) and $387 \text{ M}^{-1} \text{ min}^{-1}$ (pH 8.77). In assays with D 102 N trypsin these values were $149 \text{ M}^{-1} \text{ min}^{-1}$ (pH 7.16) and $28 \text{ M}^{-1} \text{ min}^{-1}$ (pH 8.77). The instability of TLCK and MUGB at alkaline pH values precluded these assays at higher pH values.

Ligand	Kinetic constant	Relative activity	
		Neutral pH	Alkaline pH
Z-Lys-S-Bzl	k_{cat}	4,400	18
Z-Lys-S-Bzl	k_{cat}/K_m	11,300	152
DFP	$k_{\text{obs}}/[I]$	11,300	10,000
MUGB	v_{titr}	>500	
TLCK	$k_{\text{obs}}/[I]$	5.1	1.4

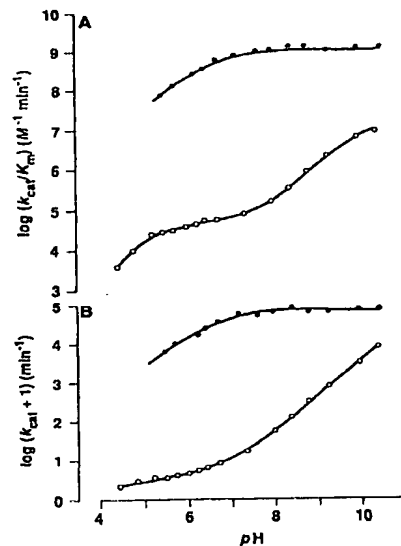
in the catalytic triad led to the postulate that Asp¹⁰² serves in concert with the histidine imidazole group to transfer the proton from the serine in a charge-relay mechanism (4). However, ¹⁵N nuclear magnetic resonance (NMR) studies (5) showed that the Asp¹⁰² and the His⁵⁷ moieties displayed normal pK_a values (K_a is the ionization constant); this is incompatible with the implications of the charge-relay mechanism (6). Furthermore, neutron diffraction and ¹H NMR studies of the imidazole nitrogens in the resting state of the enzyme show that no proton transfer occurs from His⁵⁷ to Asp¹⁰² (7). Asp¹⁰² may be involved in the stabilization of the imidazolium intermediate and the orientation of the correct tautomer of His⁵⁷ relative to Ser¹⁹⁵ and the substrate (8). However, a test of the function of Asp¹⁰² by selective chemical modification, has not been possible because it is inaccessible to chemical reagents under nondenaturing conditions. We have evaluated the catalytic role of Asp¹⁰² by replacing this residue with Asn. This eliminates the negative charge with little change in the van der Waals surface of the side-chain atoms (NH₂ versus OH).

Conversion of the Asp¹⁰² codon (GAC) to an Asn (AAC) codon within the rat anionic trypsinogen DNA (9) was accomplished by site-directed mutagenesis (10).

The DNA that encodes the mutant enzyme was sequenced in its entirety to ensure that no inadvertent base changes were introduced during the mutagenesis procedure. The mutant enzyme trypsin¹⁰² (Asp → Asn), referred to as D 102 N trypsin and the naturally occurring trypsin were expressed under the control of the simian virus 40 (SV40) early promoter (11) in stably transformed eukaryotic cell lines that secreted the zymogen form of the enzymes into the culture medium (12). D 102 N trypsin and trypsin were purified to homogeneity and crystallinity by a combination of ion-exchange and affinity chromatography techniques. Trypsin isolated from this expression system displayed physical and catalytic properties identical to trypsin purified from the rat pancreas. In contrast, D 102 N trypsin exhibited dramatically different catalytic activity.

The activities of trypsin and D 102 N trypsin toward various substrates and inhibitors are compared in Table 1. At neutral pH the catalytic efficiency of D 102 N trypsin as measured by its ability to hydrolyze the ester substrate *N*-benzyloxycarbonyl-L-lysine thiobenzyl ester (Z-Lys-S-Bzl) is severely compromised (k_{cat} or k_{cat}/K_m values are $\sim 10^4$ times lower than that of trypsin; k_{cat} is the catalytic rate constant and K_m is the

Fig. 1. Profile of activities for trypsin and D 102 N trypsin-catalyzed hydrolysis of Z-Lys-S-Bzl. (A) Plot of $\log(k_{\text{cat}}/K_m)$ versus pH and (B) plot of $\log k_{\text{cat}}$ versus pH, for trypsin (●), and D 102 N trypsin (○). Assays were performed at 25°C in 50 mM Mes [2-(*N*-morpholino)ethanesulfonic acid], Mops, or Taps buffers, pH 4.43 to 8.77, or 50 mM glycine, pH 9.25 to 10.18, that contained 0.1M NaCl and 1 mM CaCl₂. Stock solutions of Z-Lys-S-Bzl and 4,4'-dithiodipyridine were prepared in water and dimethylformamide, respectively. The pH of all reactions was determined immediately after reaction. To a cuvette that contained 0.97 ml of the assay solution was added 10 μl of a 25 mM solution of 4,4'-dithiodipyridine (final concentrations: 250 μM 4,4'-dithiodipyridine and 1% dimethylformamide) and 10 μl of a Z-Lys-S-Bzl stock solution. The concentration of substrate ranged from ten times greater than to ten times less than the K_m of the enzyme. After the background rate of hydrolysis was measured spectrophotometrically (Beckman DU-7) at 324 nm, 10 μl of an enzyme stock solution (in the case of trypsin, diluted to 0.5 mg per milliliter of bovine serum albumin) was added and the initial rate of hydrolysis was measured. At pH values greater than 9.25, for which the background hydrolysis was substantial (up to 2% Z-Lys-S-Bzl hydrolyzed per minute), a reference cell that contained substrate and 4,4'-dithiodipyridine was used during kinetic measurements. In all of the assays the initial rates were measured from data for the initial 5 to 10% of the hydrolysis of substrate. Z-Arg-S-Bzl was not used as substrate because this compound shows a background hydrolysis rate 20 times greater than that for Z-Lys-S-Bzl at alkaline pH (14). Substrate and enzyme concentration determinations were performed with standard procedures (29, 30). Values for k_{cat} and K_m parameters from all assays were derived by a program that performed a weighted linear and nonlinear squares regression analysis of data by using the Lineweaver-Burk and Michaelis-Menton equations, respectively (31). Double reciprocal plots of the data were linear in all cases. Values of pK_a and k_{cat} were determined by the program MULTI (32) which performs a nonlinear squares analysis of the data.



Michaelis constant). However, the relative activity of the mutant enzyme progressively increases with increasing pH values. To determine the relative reactivity of Ser¹⁹⁵ and His⁵⁷ both enzymes were treated with the specific active site-directed reagents DFP and TLCK. The inhibition of D 102 N trypsin by DFP, which is specific for Ser¹⁹⁵, is approximately four orders of magnitude slower than that of trypsin at both pH 8.0 and pH 10.0. The active site titrant 4-methylumbelliferyl-*p*-guanidinobenzoate (MUGB) (13) also reacts with D 102 N trypsin at a rate at least 500-fold slower than with trypsin at pH 7.5. These data suggest that the nucleophilicity of Ser¹⁹⁵ is dependent on the negative charge of Asp¹⁰².

The substrate analog TLCK reacts specifically with His⁵⁷, presumably because the binding pocket of the substrate positions the reactive chloromethyl-ketone group adjacent to His⁵⁷. In contrast to the large decreases in activity monitored with DFP and MUGB, TLCK is five times less reactive with D 102 N trypsin than with trypsin at neutral pH (pH 7.2) and one and a half times less reactive at more alkaline pH (pH 8.8). Thus the active site reacts virtually normally with the affinity reagent. The differential effect of the Asp to Asn substitution on the inhibition of D 102 N trypsin by DFP and TLCK may be due to differences in the proximity of the reactive groups of the inhibitors and the enzyme. However, a more likely explanation is that the imidazole of His⁵⁷ in D 102 N trypsin is not in the correct tautomeric state for removal of the Ser¹⁹⁵ proton and thereby reduces the reactivity of the enzyme to DFP. However, His⁵⁷ can still react with the chloromethyl ketone moiety of TLCK and thereby inhibit the enzyme.

The modified and unmodified enzymes exhibit different pH activity profiles for the ester substrate (Table 1 and Fig. 1). Similar data have been obtained with peptide substrates (14). In agreement with studies on bovine cationic trypsin (15), rat anionic trypsin shows a sigmoidal dependence of activity ($pK_a = 6.8$) with maximal k_{cat} and k_{cat}/K_m values of $7498 \pm 254 \text{ min}^{-1}$ and $1.20 \pm 0.28 \times 10^9 \text{ M}^{-1} \text{ min}^{-1}$, respectively (16, 17). The rat enzyme resembles porcine elastase (18) but differs from bovine trypsin in being alkaline stable. The dominant effect of the Asp to Asn mutation is on k_{cat} . The K_m values of the two enzymes are similar at any given pH value. The D 102 N trypsin activity is dramatically lower ($\sim 10^4$ times as measured by k_{cat} or k_{cat}/K_m) than trypsin activity at neutral pH values; however, it increases progressively at alkaline pH values from the low value at neutral pH to values

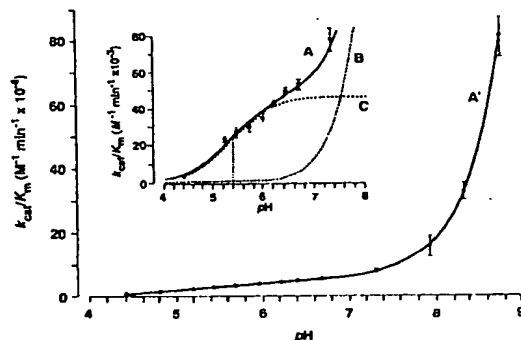


Fig. 2. The pH dependence of the kinetic parameter k_{cat}/K_m of D 102 N trypsin-catalyzed hydrolysis of Z-Lys-S-Bzl. The points correspond to the experimentally derived k_{cat}/K_m values. Curve A' is derived from substituting the calculated rate and equilibrium constants k_{OH} , k_{enz} , K_1 , and K_2 into Eq. 1. Values for k_{OH} and k_2 were determined from assays performed from pH 8.36 to 10.18 where it is assumed that $K_1 \gg [H^+]$ and $k_{OH}[OH^-] \gg k_{enz}$. Equation 1 can then be simplified and rearranged to describe a straight line: $(k_{cat}/K_m)[H^+]$

$= -K_2(k_{cat}/K_m) + (10^{-14})k_{OH}$. Linear regression of this line yields k_{OH} and K_2 values of $1.45 \pm 0.12 \times 10^{11} \text{ M}^{-2} \text{ min}^{-1}$ and $1.21 \pm 0.30 \times 10^{-10} \text{ M}$, respectively. Values of K_1 and k_{enz} were determined from assays performed from pH 4.43 to 7.33 where $[H^+] \gg K_2$. By using the k_{OH} value determined above, Eq. 1 can again be simplified to a linear form: $[k_{cat}/K_m][H^+] - 1.45 \times 10^{-3}/[H^+] = 1/K_1[1.45 \times 10^{-3} - (k_{cat}/K_m)[H^+]] + k_1$. Linear regression analysis of this line yields k_{enz} and K_1 values of $4.78 \pm 0.22 \times 10^6 \text{ min}^{-1}$ and $3.67 \pm 0.32 \times 10^{-6} \text{ M}$, respectively. Inset: Plot of k_{cat}/K_m versus pH from pH 4.43 to 7.33. Curve A is the same as described above. Curve B describes the contribution to the catalytic rate of D 102 N trypsin that depends on $[OH^-]$: $k_{OH}[OH^-]/(1 + K_2/[H^+])$. Curve C describes the contribution to the catalytic rate of D 102 N trypsin independent of $[OH^-]$ detected at lower pH values: $k_1/(1 + ([H^+]/K_1) + (K_2/[H^+]))$. Note that curve A is the sum of curves B and C. The dotted line perpendicular to the abscissa is the pK_a of the mutant enzyme calculated from the inflection point of the activity profile.

Table 2. Values for k_{OH} , k_{enz} , and pK_a derived from the D 102 N trypsin-catalyzed hydrolysis of Z-Lys-S-Bzl. The k_{OH} , k_{enz} , and pK_a parameters derived from k_{cat}/K_m values were determined as described in the legend to Fig. 2. The pK_2 values for k_2 and k_3 were not determined due to experimental constraints described below. The k_{cat} parameter does not appear to depend on the ionization of a residue in the pH range between 4 and 8. Equation 1 can then be reduced to:

$$k_{cat} = [k_{enz}/1 + (K_2/[H^+])] + [k_{OH}[OH^-]/1 + (K_2/[H^+])]$$

Values for k_{OH} and K_2 can be determined from assays performed at pH values of 8 and greater where it is assumed that $k_{OH}[OH^-] \gg k_{enz}$. The equation can then be rearranged to the linear form $k_{cat}[H^+] = -K_2k_{cat} + (10^{-14})k_{OH}$. Linear regression analysis of this line with data from assays performed from pH 7.96 to 10.18 yields a k_{OH} value of $5.50 \pm 0.21 \times 10^6 \text{ M}^{-1} \text{ min}^{-1}$ and a K_2 value of $5.89 \pm 0.50 \times 10^{-11} \text{ M}$. The value of k_{enz} can be estimated from assays performed at pH values less than 8 where $[H^+] \gg K_2$. By using the k_{OH} value determined above the equation can be reduced to $k_{enz} = k_{cat} - 5.50 \times 10^6 [OH^-]$. Subtracting the calculated $5.50 \times 10^6 [OH^-]$ values from the experimentally derived k_{cat} values from pH 4.43 to pH 7.33 gives a k_{enz} value of $0.37 \pm 0.09 \text{ min}^{-1}$. The pH dependence of the acylation rate constant k_2 of the D 102 N trypsin-catalyzed hydrolysis of Z-Lys-S-Bzl was determined by performing assays at 25°C in 50 mM Mes, Mops, or Taps buffers, pH 4.81 to 8.36 under identical conditions as for assays described in the legend to Fig. 1 except that D 102 N trypsin concentrations (4 to 40 μM) were in large excess over the initial substrate concentration (0.54 μM) and the reaction was allowed to proceed to completion. Assays performed at pH values above pH 8.4 were too fast to follow spectrophotometrically thereby preventing the determination of k_2 (acylation) values. Values for k_2 and K_m were determined by the procedure of Keszdy and Bender (28). The k_{OH} parameter was obtained from a plot of the k_2 values versus solvent hydroxide ion concentration from pH 6.70 to 8.36; $k_{OH} = 4.91 \pm 0.72 \times 10^6 \text{ M}^{-1} \text{ min}^{-1}$. Values for k_{enz} and K_1 were obtained by using the k_{OH} value of $4.91 \times 10^6 \text{ M}^{-1} \text{ min}^{-1}$ and by rearranging Eq. 1 with $[H^+] \gg K_2$ to yield:

$$(k_2[H^+] - 4.91 \times 10^{-8})/[H^+] = (1/K_1)(4.91 \times 10^{-8} - k_2[H^+]) + k_{enz}$$

Linear regression analysis of this line with k_2 values determined from assays performed from pH 4.81 to pH 6.70 yielded a k_{enz} value of $1.32 \pm 0.08 \text{ min}^{-1}$ and a K_1 value of $5.35 \pm 1.00 \times 10^{-6} \text{ M}$. Values for k_3 (deacylation) were calculated using the experimentally derived k_{cat} and k_2 values and the equation: $k_3 = (k_{cat}k_2)/(k_2 - k_{cat})$. The k_{OH} value was determined from a plot of the k_3 values versus solvent hydroxide ion concentration from pH 6.70 to 8.36; $k_{OH} = 4.57 \pm 2.43 \times 10^7 \text{ M}^{-1} \text{ min}^{-1}$. The maximal value of the deacylation rate constant of the hydroxide-independent pathway, k_{enz} , was calculated by incorporating the k_{enz} values for k_2 and k_{cat} determined above into the equation $k_3 = k_{cat}k_2/(k_2 - k_{cat})$. This gives a k_{enz} (deacylation) of $0.51 \pm 0.07 \text{ min}^{-1}$. The value of k_3 like k_{cat} shows no dependence on the ionization of a residue in the pH range between 5 and 8.

Rate constant	k_{OH} ($\text{M}^{-1} \text{ min}^{-1}$)	k_{enz} (min^{-1})	pK_1	pK_2
k_{cat}	5.50×10^6	0.37		10.2
k_{cat}/K_m	$1.45 \times 10^{11} \text{ M}^{-1}$	$4.78 \times 10^6 \text{ M}^{-1}$	5.4	9.9
k_2	4.91×10^6	1.32	5.3	
k_3	4.17×10^7	0.51		

that approach those of the native enzyme (k_{cat}/K_m 6% at pH 10.2).

The ascendant alkaline limb of the activity-pH profiles of the D 102 N trypsin is not an artifact due to deamidation of the Asn residue to Asp, since mutant enzyme activity at neutral pHs is not affected by preincubation at alkaline pH. Furthermore, one would expect the pH activity profiles to be similar in shape to those of the naturally occurring enzyme if they merely reflected contamination by trypsin. We ascribe this ascendant basic limb to the participation of a titratable base or bases or of OH⁻ itself. Although the mechanism of catalysis by the D 102 N trypsin is unknown, the pH rate profile of k_{cat}/K_m can be described by a bipartite rate equation in which one part represents the catalytic rate detected at the lower pH values and the other part describes the catalytic rate that shows a dependence on hydroxide ion concentration (19). The observed rate constant k_{cat}/K_m can be defined as:

$$k_{cat}/K_m = \frac{k_{enz}}{1 + ([H^+]/K_1) + (K_2/[H^+])} + \frac{k_{OH}[OH^-]}{1 + (K_2/[H^+])} \quad (1)$$

where k_{enz} is the rate constant of the hydroxide independent pathway, K_1 and K_2 are the dissociation constants of the ionizing groups, and k_{OH} is the rate constant of the hydroxide ion dependent pathway. The catalytic activity of the OH⁻-activated and OH⁻-independent pathways can be resolved with Eq. 1. Values for k_{cat}/K_m determined from mutant enzyme activity studies above pH 8.0 show an increase with solvent hydroxide ion concentration that yields k_{OH} and K_2 values of $1.45 \pm 0.12 \times 10^{11} M^{-2} \text{ min}^{-1}$ and $1.21 \pm 0.30 \times 10^{-10} M$ ($pK_2 = 9.9$), respectively. Between pH 8.0 and pH 8.8 the k_{cat}/K_m values increase linearly with hydroxide ion concentration. The slight decrease from linearity above pH 8.8 may reflect the ionization of another group with an alkaline pK_a value such as the lysine substrate or the amino-terminal group of the protein (20).

There is good agreement between the calculated k_{cat}/K_m curve derived from Eq. 1 and the experimentally derived values (Table 2 and Fig. 2). Measurements of k_{cat}/K_m values below pH 8.0 yield k_{enz} and K_1 values of $4.78 \pm 0.22 \times 10^6 M^{-1} \text{ min}^{-1}$ and $3.67 \pm 0.32 \times 10^{-6} M$ ($pK_1 = 5.4$), respectively. A comparison of the k_{enz} value for D 102 N trypsin and the maximal k_{cat}/K_m value for trypsin indicates that the activity of the mutant enzyme (ignoring the contribution of the OH⁻-dependent pathway) is 25,000 times less than that of trypsin. Thus Asp¹⁰² is crucial for the catalytic activity at neutral

pH values. However, the rate of hydrolysis by the mutant enzyme is still 400 times greater than the rate of solvent hydrolysis of the substrate. The inflection points of the curves in Fig. 2 suggests that the pK_a of His⁵⁷ has decreased 1.5 pH units in D 102 N trypsin compared to trypsin. The putative alteration in the pK_a value of His⁵⁷ reflects the replacement of the negatively charged carboxylate group with a neutral amide group. The mutant enzyme exhibits classic burst kinetics on ester substrates below pH 7.0. This implies that an acyl enzyme intermediate accumulates and that deacylation is rate determining in this pH range (14).

It has been suggested that Asp¹⁰² controls the position of the neighboring His⁵⁷ residue that in turn modulates the polarity of the Ser¹⁹⁵ (8). Our demonstration of the crucial role of Asp¹⁰² is not surprising in view of the strict evolutionary conservation of this residue within the catalytic triad. The magnitude of the catalytic defect from the Asp¹⁰² → Asn replacement and the alkaline activation of the enzyme are unexpected. The three-dimensional structure of D 102 N trypsin is virtually identical to that of trypsin in the alkaline pH range (21). Thus the activity of the mutant enzyme arises from an active site conformation that resembles the native structure. Certain properties of the D 102 N trypsin superficially resemble chymotrypsin methylated at His⁵⁷ (22). The activity of both enzymes is dramatically lower at neutral pH values and increases in proportion to OH⁻ concentration. However, the rate constant ascribed to the reaction with OH⁻ ions is 1000 times greater for the D 102 N trypsin mutant than for chymotrypsin with the modified histidine. Nevertheless, these results are consistent with the view that compromising the function of the histidine dramatically decreases catalytic activity at neutral pH values. This defect can be partly overcome at basic pH. The alkaline pH may affect the catalytic reaction indirectly by affecting the ionization of groups that function in catalysis. Alternatively, OH⁻ might participate directly in the reaction; this would require activation at very low hydroxide ion concentrations. The overall catalytic mechanism of the D 102 N trypsin activity is unknown at present. The activity may be due in part to a nucleophilic contribution from the imidazole nitrogen of His⁵⁷ instead of Ser¹⁹⁵ as has been detected in the cleavage of active esters of nonspecific substrates (23). Alternatively, a residue distant from the active site may contribute to stabilization of the tetrahedral intermediate at basic pH. Whatever the mechanism of action, D 102 N trypsin displays distinctive properties that distinguish it from trypsin. Its low activity in the neutral pH range

makes it an unattractive catalyst for most biological functions; thus it might not be expected to persist in evolution. The Asn mutant, however, is of considerable interest as a distinctive serine protease. This work illustrates the potential for creating new variants that are not found in nature because they are active under extreme conditions that are usually incompatible with cellular environments.

REFERENCES AND NOTES

1. J. J. Birktoft and D. M. Blow, *J. Mol. Biol.* 68, 187 (1972); A. Tulinsky et al., *Biochemistry* 12, 4185 (1973); R. M. Stroud et al., *J. Mol. Biol.* 83, 185 (1974); R. Huber et al., *ibid.* 89, 73 (1974); L. Sawyer et al., *ibid.* 118, 137 (1978); W. Bode et al., *ibid.* 164, 237 (1983); C. S. Wright et al., *Nature (London)* 221, 235 (1969); G. D. Brayer et al., *J. Mol. Biol.* 124, 261 (1978); P. W. Coddling et al., *Can. J. Biochem.* 52, 208 (1974); G. D. Brayer et al., *J. Mol. Biol.* 131, 743 (1979).
2. G. H. Dixon, S. Go, H. Neurath, *Biochim. Biophys. Acta* 19, 193 (1956).
3. E. Shaw, M. Mares-Guia, W. Cohen, *Biochemistry* 4, 2219 (1965).
4. D. M. Blow, J. J. Birktoft, B. S. Hartley, *Nature (London)* 221, 337 (1969).
5. W. W. Bachovchin and J. D. Roberts, *J. Am. Chem. Soc.* 100, 8041 (1978).
6. G. A. Rogers and T. C. Bruice, *ibid.* 96, 2473 (1974).
7. A. A. Kossiakoff and S. A. Spencer, *Nature (London)* 288, 414 (1980); J. L. Markley and I. B. Ibanez, *Biochemistry* 22, 4627 (1978).
8. L. Polgar and M. L. Bender, *Proc. Natl. Acad. Sci. U.S.A.* 64, 1335 (1969); A. R. Fersht and J. Sperling, *J. Mol. Biol.* 74, 137 (1973).
9. C. S. Craik et al., *Science* 228, 291 (1985).
10. M. J. Zoller and M. Smith, *DNA* 3, 479 (1984).
11. P. J. Southern and P. Berg, *J. Mol. App. Genet.* 4, 327 (1982).
12. This was accomplished as follows: Chinese hamster ovary cells were co-transfected with a plasmid that contained either the trypsinogen or D 102 N trypsinogen DNA constructs under transcriptional control of the T-antigen early promoter of SV40 and a plasmid that encoded the bacterial phosphotransferase gene (*neo*). The *neo* gene conferred resistance to the amino-glycoside antibiotic G418 and permitted the phenotypic selection of a cell line with high probability of co-expressing the trypsinogen gene. A filter screening assay was developed for detecting high levels of protein secretion from transfected cells in order to isolate cell lines that overproduced trypsinogen (C. S. Craik and R. L. Burke, unpublished results). Cell lines that produced trypsinogens in large amounts (about 10 mg/liter) were then expanded into mass culture (40 liters).
13. J. C. McRae et al., *Biochemistry* 20, 7196 (1981).
14. C. S. Craik et al., unpublished results.
15. T. Inagami and J. M. Sturtevant, *Biochim. Biophys. Acta* 38, 64 (1960); H. P. Kasserra and K. J. Laidler, *Can. J. Chem.* 47, 4021 (1969).
16. The catalysis constant k_{cat} is used as a measure of catalytic activity and is a first-order rate constant that refers to the properties and reactions of the enzyme-substrate, enzyme-intermediate, and enzyme-product complexes. The Michaelis constant K_m relates to the binding affinity of the enzyme for its substrate and is an apparent dissociation constant that may be treated as the overall dissociation constant of all enzyme-bound species. The ratio of k_{cat}/K_m is an apparent second-order rate constant that refers to the properties and reactions of the free enzyme and free substrate (17).
17. A. Fersht, *Enzyme Structure and Mechanism* (Freeman, New York, 1985).
18. P. Geneste and M. L. Bender, *Proc. Natl. Acad. Sci. U.S.A.* 64, 683 (1969).
19. The equation fits the experimental points and suggests a mechanism that involves a combination of an acid HA and OH⁻. However, by the principle of kinetic equivalence it cannot be distinguished from a

- mechanism involving A⁺.
20. M. L. Bender *et al.*, *J. Am. Chem. Soc.* 86, 3680 (1964); A. Himoe and G. P. Hess, *Biochem. Biophys. Res. Commun.* 23, 234 (1966).
 21. S. Sprang *et al.*, *Science* 237, 905 (1987).
 22. R. Henderson, *Biochem. J.* 124, 13 (1971); J. Faure and N. Houyet, *Eur. J. Biochem.* 81, 515 (1977).
 23. C. D. Hubbard and J. F. Kirsch, *Biochemistry* 11, 2483 (1972).
 24. R. Kitz and F. S. Wilson, *J. Biol. Chem.* 237, 3245 (1962).
 25. Incubations with diisopropylfluorophosphate (DFP) were performed at 25°C with 100 nM wild-type or mutant enzyme and varying concentrations of DFP in either 50 mM Tris [3-[[tris(hydroxymethyl)methyl]amino] propanesulfonic acid], pH 7.96, or 50 mM glycine, pH 10.03, that contained 0.1 M NaCl, 1 mM CaCl₂, 0.005% (w/v) Triton X-100 and 5% (v/v) isopropanol. DFP stock solutions were made up in isopropanol. Final volumes were 0.200 ml and 6.2 ml when incubations were performed with trypsin and D 102 N trypsin, respectively. Trypsin enzyme activities were measured spectrophotometrically at 324 nm by adding 10 µl of the trypsin-DFP solution to 0.99 ml of the same buffer (1 nM trypsin final concentration) that contained 60 µM *N*-benzyloxycarbonyl-L-lysine benzylthioester (Z-Lys-S-Bzl) and 250 µM 4,4'-dithiodipyridine but no DFP. Mutant enzyme activities at 324 nm were determined by adding 10 µl of 6 mM Z-Lys-S-Bzl and 10 µl of 25 mM 4,4'-dithiodipyridine to 0.98 ml of the D 102 N trypsin-DFP solution. The concentrations of DFP during incubations with trypsin at both pH values were 0, 20, 25, 40, 80, or 200 µM. In incubations with D 102 N trypsin initial DFP concentrations were 0, 5, 8, 10, or 12.5 mM, and 0, 10, 12.5, or 16.6 mM when assays were performed at pH 7.96 and pH 10.03, respectively.
 26. T. Yoshimura, L. N. Barker, J. C. Powers, *J. Biol. Chem.* 257, 5077 (1982).
 27. Incubations with tosyl-L-lysine chloromethyl ketone (TLCK) were performed at 25°C with 5 µM unmodified or mutant enzyme in 100 µl of either 100 mM Mops [3-*N*-(morpholino)propanesulfonic acid], pH 7.16 or 100 mM Tris, pH 8.77, that contained either 0 or 200 µM TLCK. Immediately before assaying trypsin activity an aliquot from the incubation mixtures was diluted 20-fold in 50 mM incubation buffer that contained 0.1 M NaCl, 1 mM CaCl₂, and 0.005% (w/v) Triton X-100. Ten microliters of the diluted enzyme solution was added to 0.99 ml of the dilution buffer that contained 250 µM 4,4'-dithiodipyridine and 100 µM Z-Lys-S-Bzl (assay buffer); enzyme activities were followed at 25°C spectrophotometrically at 324 nm (2.5 nM trypsin). Mutant enzyme activities were determined in an identical manner as described above except that 10 µl of the incubation mixture that contained D 102 N trypsin was added directly to 0.99 ml of the assay buffer (50 nM D 102 N trypsin). After 3 hours of incubation with TLCK the loss of catalytic activity of both enzymes at both pH values was complete. Since there was no activity of either enzyme after 2.5 hours of further incubation with a large excess of substrate over TLCK, the inhibition was probably due to formation of a covalent bond between the enzyme and the inhibitor and not to a slow-off rate for a noncovalently bound competitive inhibitor.
 28. F. J. Kezdy and M. L. Bender, *Biochemistry* 1, 1097 (1962).
 29. Substrate concentrations were calculated by using the total change in absorbance at 324 nm when reactions were run at pH 7.5 ($\epsilon_{324} = 19,800 \text{ M}^{-1} \text{ cm}^{-1}$ (13)) and catalyzed by 1 nM trypsin so that the reaction would be completed within several minutes. The molar absorptivities over the pH range of 4.1 to 10.6 were determined by repeating these reactions with known substrate concentrations at various pH values. It was found that at pH values below 7.6 the molar absorptivity remained at $19,800 \text{ M}^{-1} \text{ cm}^{-1}$, but that this value decreased sigmoidally above pH 7.6, with an apparent pK_a of 8.7, presumably reflecting the ionization of thiopyridine. These molar absorptivity values were used to convert rates of reaction into molar rates at the appropriate pH values.
- The concentration of viable active sites in native enzyme preparations were determined by active site titrations with 4-methylumbelliferyl *p*-guanidinobenzoate (MUGB) by using 4-methylumbelliferone as a standard (30) as described in the legend to Table 1. Titrations of D 102 N trypsin proved to be too slow to measure active sites accurately. Mutant enzyme concentrations were thus determined in duplicate by absorbance at 280 nm ($\epsilon_{280} = 38,000 \text{ M}^{-1} \text{ cm}^{-1}$). The accuracy of this molar absorptivity value was confirmed by amino acid analysis with norleucine as an internal standard.
- A danger in following the activity of D 102 N trypsin is that an unknown proportion of the activity may be due to trypsin that has formed through deamidation. This does not appear to be a problem at pH values less than 8 where the activity of the mutant enzyme is less than 0.1% that of trypsin. At alkaline pH values, where the activity of the mutant enzyme becomes significant, the possibility of activity resulting from deamidation becomes greater. However, assays with 100 nM D 102 N trypsin and 60 µM Z-Lys-S-Bzl as substrate at pH 7.16 and pH 10.24 after prior incubation of the enzyme in buffers at either pH value for 1 hour gave initial rates of reaction of $1.00 \pm 0.04 \text{ min}^{-1}$ and $249 \pm 5 \text{ min}^{-1}$, respectively. These results indicate that significant deamidation of the D 102 N residue to an aspartic acid did not occur in the pH and time ranges studied.
30. G. W. Jameson, D. V. Roberts, R. W. Adams, S. A. Kyle, D. T. Elnore, *Biochem. J.* 131, 107 (1973).
 31. D. V. Roberts, in *Enzyme Kinetics* (Cambridge Univ. Press, Cambridge, 1977), p. 299.
 32. K. Yamaoka *et al.*, *J. Pharm. Dyn.* 4, 879 (1981).
 33. We thank J. F. Kirsch and E. T. Kaiser for helpful discussions and L. Spector for preparing the manuscript. Support by NSF grant PCM830610 (W.J.R.) and DMB8608086 (C.S.C.), and a Bristol Meyer grant of Research Corporation (C.S.C.) is gratefully acknowledged. An NIH postdoctoral fellowship was awarded to S.R. (GM 10765).

29 September 1986; accepted 29 May 1987

Adrenal Medulla Grafts Enhance Recovery of Striatal Dopaminergic Fibers

MARTHA C. BOHN,* LISA CUPIT, FREDERICK MARCIANO, DON M. GASH

The drug, 1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine (MPTP), depletes striatal dopamine levels in primates and certain rodents, including mice, and produces parkinsonian-like symptoms in humans and nonhuman primates. To investigate the consequences of grafting adrenal medullary tissue into the brain of a rodent model of Parkinson's disease, a piece of adult mouse adrenal medulla was grafted unilaterally into mouse striatum 1 week after MPTP treatment. This MPTP treatment resulted in the virtual disappearance of tyrosine hydroxylase-immunoreactive fibers and severely depleted striatal dopamine levels. At 2, 4, and 6 weeks after grafting, dense tyrosine hydroxylase-immunoreactive fibers were observed in the grafted striatum, while only sparse fibers were seen in the contralateral striatum. In all cases, tyrosine hydroxylase-immunoreactive fibers appeared to be from the host rather than from the grafts, which survived poorly. These observations suggest that, in mice, adrenal medullary grafts exert a neurotrophic action in the host brain to enhance recovery of dopaminergic neurons. This effect may be relevant to the symptomatic recovery in Parkinson's disease patients who have received adrenal medullary grafts.

IN HUMANS, THE DRUG, 1-METHYL-4-PHENYL-1,2,5,6-TETRAHYDROPYRIDINE (MPTP), produces motor deficits that closely resemble those observed in Parkinson's disease (1-4). This observation has led to the development of animal models of Parkinson's disease that are valuable for studying the effects of brain grafting (5). MPTP damages the dopamine (DA)-containing A9 cell group in the pars compacta of the substantia nigra and results in a degeneration of the nigrostriatal DA fibers and loss of striatal DA and its metabolites (1-8). The severity of this damage is species-dependent. In primates, MPTP treatment damages both the DA fibers and cell bodies (1-5). In mice, the fibers are damaged, but many A9 neurons survive (6, 7). Because the MPTP lesion is transient in mouse (7, 9), the MPTP-treated mouse provides an opportunity for studying recovery of identified neurons in the brain. Our study suggests

that striatal grafts of adult mouse adrenal medulla enhance recovery of these neurons.

Two MPTP treatments were compared for their effects on striatal DA levels and tyrosine hydroxylase-immunoreactivity (TH-IR) in the striatum and A9 region of C57BL/6 mice (6 to 12 weeks old; 21 to 28 g). As described (6, 7), lightly etherized mice received multiple injections of MPTP-HCl subcutaneously in 0.5 ml of saline. Group A received three injections of 30 mg per kilogram of body weight at 24-hour intervals and group B received two injections of 50 mg per kilogram of body weight 16 hours apart. Catecholamines in tissues were isolated and measured

M. C. Bohn and L. Cupit, Department of Neurobiology and Behavior, State University of New York, Stony Brook, NY 11794.

F. Marciano and D. M. Gash, Department of Neurobiology and Anatomy, University of Rochester School of Medicine, Rochester, NY 14642.

*To whom correspondence should be addressed.

Localization of the mosaic transmembrane serine protease corin to heart myocytes

John D. Hooper¹, Anthony L. Scarman¹, Belinda E. Clarke², John F. Normyle¹ and Toni M. Antalis¹

¹Cellular Oncology Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia;

²Department of Anatomical Pathology, The Prince Charles Hospital, Chermside, Queensland, Australia

Corin cDNA encodes an unusual mosaic type II transmembrane serine protease, which possesses, in addition to a trypsin-like serine protease domain, two frizzled domains, eight low-density lipoprotein (LDL) receptor domains, a scavenger receptor domain, as well as an intracellular cytoplasmic domain. In *in vitro* experiments, recombinant human corin has recently been shown to activate pro-atrial natriuretic peptide (ANP), a cardiac hormone essential for the regulation of blood pressure. Here we report the first characterization of corin protein expression in heart tissue. We generated antibodies to two different peptides derived from unique regions of the corin polypeptide, which detected immunoreactive corin protein of approximately 125–135 kDa in lysates from human heart tissues. Immunostaining of sections of human heart showed corin expression was specifically localized to the cross striations of cardiac myocytes, with a pattern of expression consistent with an integral membrane localization. Corin was not detected in sections of skeletal or smooth muscle. Corin has been suggested to be a candidate gene for the rare congenital heart disease, total anomalous pulmonary venous return (TAPVR) as the corin gene colocalizes to the TAPVR locus on human chromosome 4. However examination of corin protein expression in TAPVR heart tissue did not show evidence of abnormal corin expression. The demonstrated corin protein expression by heart myocytes supports its proposed role as the pro-ANP convertase, and thus a potentially critical mediator of major cardiovascular diseases including hypertension and congestive heart failure.

Keywords: serine protease; corin; heart; pro-atrial natriuretic peptide (pro-ANP); TAPVR.

Serine proteases are found in all living organisms, ranging from viruses to humans [1], where they serve important and varied biological functions in situations requiring limited proteolysis. Their activities impact on areas as diverse as hemostasis, tissue remodelling and wound repair, inflammation, angiogenesis, fibrinogenesis and fibrinolysis. Cell surface serine proteases have been associated largely with extracellular matrix degradation, but there are emerging roles for these proteases in generating bioactive matrix protein fragments, influencing the release, the activation and bioavailability of growth factors and in shedding of cell surface proteins [2–6].

Many serine proteases are mosaic proteins comprising multiple, structurally distinct domains necessary for regulating enzymatic activity. Circulating serine proteases of the blood coagulation (e.g. prothrombin and factor X) [7], fibrinolysis (e.g. plasminogen activators) [8] and complement (e.g. C1r and C1s) [9] systems are well characterized examples of mosaic proteins. While the vast majority of known serine proteases are secreted, more recently some serine proteases have been found to possess integral transmembrane domains. The proteins enteropeptidase [10], hepsin [11] and most recently, TMPRSS2

[12] are examples of mosaic serine proteases with type II transmembrane domains. These enzymes are positioned on the plasma membrane via a membrane spanning domain close to the N-terminus. In addition to membrane spanning and protease domains, enteropeptidase also contains two low-density lipoprotein (LDL) receptor domains, a meprin-like domain, two C1r-like domains and a truncated scavenger receptor domain. An LDL receptor domain and a scavenger receptor domain have also been identified in TMPRSS2 [12]. The functions of these domains have not been determined.

Serine proteases play important roles in several aspects of heart physiology and cardiovascular disease [13]. The mast cell serine protease chymase is believed to be the major converter of angiotensin (ang)I to angII in human heart tissue [14]. The involvement of angII in normal cardiac function as well as in heart ailments such as hypertrophy, heart failure and ischaemic heart disease is indicated by the finding that inhibition of the angiotensin converting enzyme (ACE), leads to beneficial outcomes for sufferers of these diseases [15]. However, ACE inhibitors block only 10–20% of angI conversion in heart tissue whereas the remaining activity is blocked by serine protease inhibitors [16]. The fibrinolytic serine proteases tissue-type plasminogen activator (tPA) and urokinase-type plasminogen activator (uPA) are also thought to be involved in the progression of heart disease. uPA is present at significantly elevated levels in the atherosclerotic lesions responsible for myocardial infarction and failure [17]. The reduction in tPA from arteriolar smooth muscle cells is linked to the development of coronary artery disease in transplanted hearts [18].

Our own work and that of Yan *et al.* [19] has led to the recent cloning of a cDNA encoding a novel, multidomain type II transmembrane serine protease from human heart. The

Correspondence to T. M. Antalis, Queensland Institute of Medical Research, Post Office Royal Brisbane Hospital, Brisbane, 4029, Queensland, Australia. Fax: + 61 73362 0107, Tel.: + 61 73362 0312, E-mail: toniA@qimr.edu.au

Abbreviations: LDL, low-density lipoprotein; ANP, atrial natriuretic peptide; TAPVR, total anomalous pulmonary venous return; tPA, tissue-type plasminogen activator; uPA, urokinase-type plasminogen activator; ang, angiotensin; ACE, angiotensin converting enzyme.

(Received 24 July 2000, revised 12 September 2000, accepted 4 October 2000)

predicted protein, corin, comprises two frizzled domains, eight LDL receptor domains, a truncated scavenger receptor domain, in addition to the extracellular trypsin-like serine protease domain [19]. Recent expression of recombinant corin demonstrates that it possesses pro-atrial natriuretic peptide (ANP) convertase activity [20], and thus may play a critical role in the regulation of hypertension. *In situ* hybridization studies of mouse embryonic heart showed that corin mRNA was expressed as early as day 9.5 and maintained its expression through the adult animal [19]. The corin gene was mapped to human chromosome 4p12–13 [19], near the locus for the congenital heart disease, total anomalous pulmonary venous return (TAPVR). Here we present data describing for the first time native corin protein expression and localization in human heart.

MATERIALS AND METHODS

Identification of corin cDNA by homology cloning

Homology cloning was performed by RT-PCR using degenerate oligonucleotides corresponding to conserved regions of serine proteases [21–24]. Total RNA was isolated from S1a cells [25] following treatment with TNF α and cycloheximide for 4 h. RNA (5 μ g) was reverse transcribed at 42 °C using AMV reverse transcriptase (Promega, Madison, WI) in the presence of oligo dT_{12–18} (0.25 μ g· μ L⁻¹) (Pharmacia Biotech, Sweden), 50 mM Tris/HCl, pH 8.3, 50 mM KCl, 10 mM MgCl₂, 10 mM dithiothreitol and 0.5 mM spermidine in a total volume of 20 μ L. PCR was performed using 1 μ L of the reverse transcriptase reaction mixture, 500 ng of each primer, 10 mM Tris HCl, pH 8.3, 50 mM KCl, 1.5 mM MgCl₂, 0.2 mM dNTPs and 1–2 units of Taq polymerase (Perkin Elmer). The primers were as follows. Forward, 5'-ACAGAATTCTGGGTIGTIACI-GCIGCICAYTG-3'; reverse, 5'-ACAGAATTCAIXGGICCI-CCI(C/G)(T/A)XTCICC-3'; where X = A or G, Y = C or T; I = inosine).

Cycling conditions: 2 cycles of 94 °C for 2.5 min, 35 °C for 2.5 min and 72 °C for 3 min, followed by 33 cycles of 94 °C for 2.5 min, 57 °C for 2.5 min and 72 °C for 3 min, with a final extension at 72 °C for 7 min. PCR products of approximately 450 bp were ligated into pGEM-T (Promega, Madison, WI, USA), cloned and analysed by DNA sequencing. A DNA fragment was identified which represented the partial corin sequence (nucleotides 334–748). The cDNA was extended 333 nucleotides towards the 5' end by screening a cDNA library using two rounds of PCR and the nested oligonucleotides ATC2P3 and ATC2P1 in combination with the vector specific primer T7. The 3' end was extended to nucleotide 976 by two rounds of PCR and the nested oligonucleotides ATC2P4 and ATC2P5 in combination with the vector specific primer T3. The primer sequences are given below.

ATC2P1: 5'-GCGTGTCTGCATGAACACTG-3'; ATC2P2: 5'-ATGCCAAGCACCACCTTCCA-3'; ATC2P3: 5'-ATAGTC-CACCACTGCTCGAC-3'; ATC2P4: 5'-TTAAGCTGCAAGA-GGGAGAG-3'.

The DNA sequence of this cDNA has been deposited in the DDBJ/Genbank/EMBL database under accession no. AF113248.

Heart tissue specimens

Tissues from explanted hearts with terminal heart failure were either snap frozen in liquid nitrogen (for RNA and protein analyses) or processed for routine histological examination. Six

paraffin embedded blocks of human heart tissue were obtained from autopsy cases with acute myocardial infarction. These blocks included both viable and nonviable myocardium. Procedures were in accordance with guidelines established by the National Health and Medical Research Council of Australia, Ethics Approval number EC9876(II).

Northern and Poly(A)⁺ RNA dot blot analyses

Human multiple tissue northern blots (Clontech, Palo Alto, CA, USA) contained 2 μ g of poly(A)⁺ RNA per lane. The blots were hybridized with a ³²P-dCTP labeled *Eco*RI digested DNA fragment encoding corin cDNA in ExpressHyb (Clontech) solution at 65 °C and washed to a final stringency of 0.2 \times NaCl/Cit, 0.1% SDS at 65 °C. The blot was reprobed with β -actin as a measure of loading in each lane. For the mouse tissue blot, total RNA was purified from mouse tissues, separated by denaturing gel electrophoresis and transferred to Hybond-N nylon membranes as described [26]. The blot was hybridized with the radiolabelled human corin DNA probe under lower stringency conditions in ExpressHyb solution at 55 °C and washed to a final stringency of 1 \times NaCl/Cit, 0.1% SDS at 55 °C. The mouse tissue blot was stained with ethidium bromide to confirm RNA loading in each lane.

Production of affinity purified anti-peptide polyclonal antibodies

Rabbit polyclonal antibodies were generated against corin specific peptides derived from nonhomologous hydrophilic regions within the corin amino-acid sequence. Two peptides, each containing a cysteine residue incorporated at the C-terminus, were synthesized (Auspep, Parkville, Australia) and conjugated to keyhole limpet hemocyanin using μ -maleimidobenzoic acid *N*-hydroxysuccinimide ester. The peptides were: A1: IQEQE-KEPRWLTLHSNWE-C, A2: GHMGNKMPFKLQEGE-C. Rabbit antisera was peptide-affinity purified using SulfoLink coupling gel (Pierce, Rockville, IL). The specificity of each antibody was tested against the immunogenic peptide by ELISA.

Western blot analysis

Frozen heart tissue (100 mg) was homogenized in lysis-binding buffer (Dynabeads mRNA Direct kit, Dynal) and spun at 13000g for 2 min. The protein pellet was dissolved in reducing SDS-sample buffer for Western blot analysis. Proteins were separated by SDS/PAGE on 10% acrylamide gels and transferred electrophoretically to Hybond-P membranes (Amersham, Aylesbury, UK). Membranes were blocked with 5% nonfat skim milk powder in Tris/NaCl (10 mM Tris/HCl, pH 7.0, 150 mM NaCl), incubated with affinity purified anti-peptide antibody, then with horseradish peroxidase conjugated sheep anti-(rabbit Ig) secondary antibody, and visualized by enhanced chemiluminescence (Amersham, Aylesbury, UK).

Immunohistochemistry

Paraffin sections (5 μ m) of formalin-fixed human heart were deparaffinized, then rehydrated before antigen retrieval in boiling 10 mM citric acid buffer, pH 6. After cooling, endogenous peroxidase activity was inhibited by 10 min incubation in 1% hydrogen peroxide. Non-specific antibody binding was blocked by incubating the sections in 4% nonfat skim milk powder in NaCl/P_i for 15 min, followed by 10%

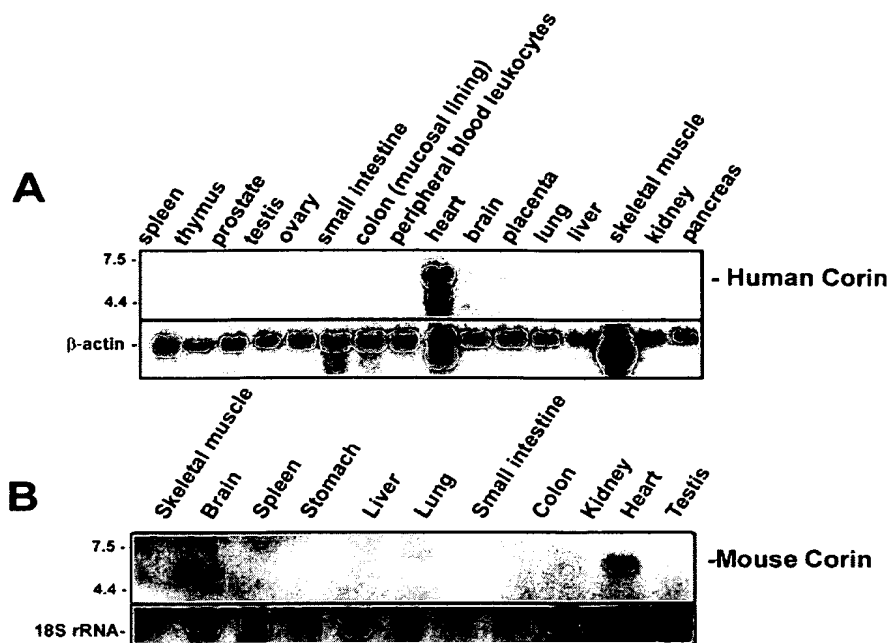


Fig. 1. Corin expression in human and mouse tissues. (A) Northern blot analysis of RNA isolated from a range of normal human tissues probed with 32 P-labelled corin cDNA. The levels of β -actin mRNA are shown as a control for loading. (B) Northern blot analysis of corin mRNA expression in a range of mouse tissues probed with 32 P-labelled human corin cDNA at reduced stringency. The levels of 18S ribosomal RNA are shown as a control for loading.

normal goat serum for 20 min. Affinity purified anticorin A1 (1 : 100; $150 \mu\text{g}\cdot\text{mL}^{-1}$) or A2 antibodies (1 : 50; $20 \mu\text{g}\cdot\text{mL}^{-1}$) were applied and incubated overnight in a humidified chamber at room temperature. Controls included sections incubated with no primary antibody or antibody that had been preadsorbed for 2 h at room temperature with $1 \mu\text{g}$ of the antigenic peptide. Following incubation with prediluted biotinylated goat anti-(rabbit Ig) Ig (Zymed, San Francisco, CA, USA), streptavidin-horseradish peroxidase (Zymed) was applied and color developed using the chromogen 3,3'-diaminobenzidine with hydrogen peroxide as substrate. The sections were counterstained in Mayers' haematoxylin.

RESULTS AND DISCUSSION

Isolation of human corin cDNA by homology cloning

A PCR-based homology cloning approach was employed to identify serine protease cDNAs expressed by the S1a cell line [25] which is resistant to tumor necrosis factor- α induced apoptosis. Degenerate primers designed to anneal to cDNA encoding the conserved regions surrounding the catalytic histidine and serine amino acids of serine proteases [21–23], were used to amplify and then clone a range of DNA fragments of approximately 450 bp. One clone, designated ATC2, was found to encode a novel serine protease. The cDNA was extended in the 5' and 3' directions by library screening and the DNA sequence was deposited in the DDBJ/Genbank/EMBL database (accession no. AF113248). This sequence was subsequently determined to be 100% identical to a recently reported cDNA encoding the serine protease, corin (accession no. AF133845) [19].

Corin mRNA is strongly expressed in heart

The tissue distribution of corin mRNA was examined by Northern blot analyses. Analysis of poly(A) $^{+}$ RNA from 16

normal human tissues showed a single transcript of approximately 5.1 kb detectable only in human heart (Fig. 1A). Examination of a range of mouse tissues also demonstrated specific expression of corin mRNA of approximately 5.1 kb only in mouse heart (Fig. 1B).

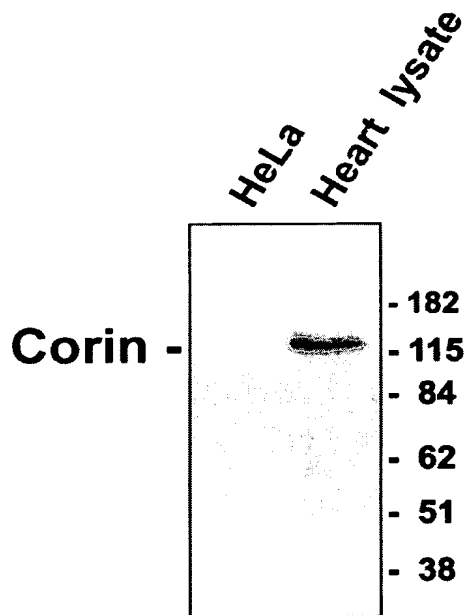


Fig. 2. Corin protein expression in human heart tissue by Western blot analysis. Immunoreactive corin protein of 125–135 kDa is detected in a protein lysate prepared from human heart tissue (Patient #7684), which is not detectable in a corin negative HeLa cell lysate. The blot was probed with anticorin antibody, AbA1, and visualized using enhanced chemiluminescence. The protein standards in kDa are as indicated.

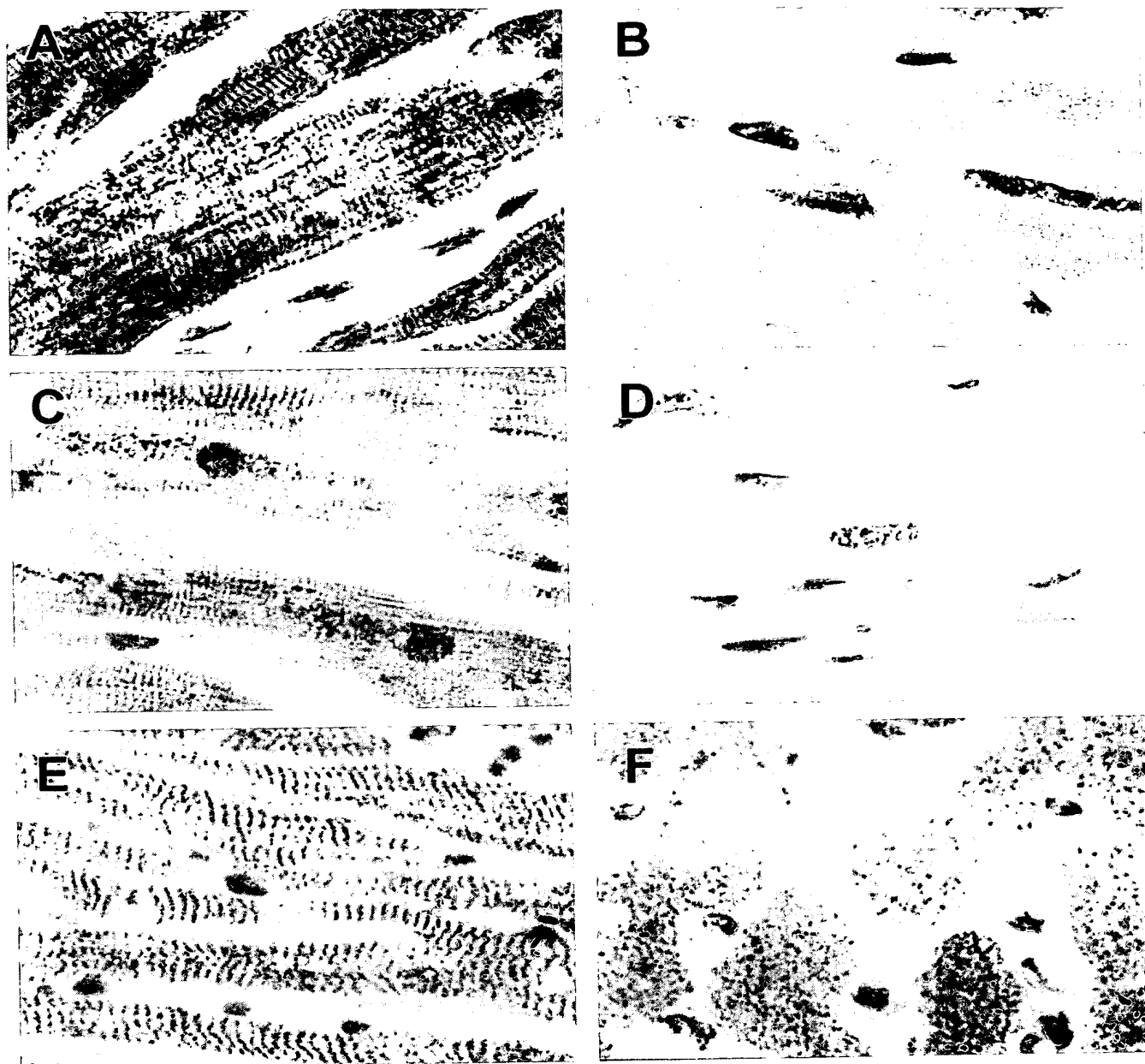


Fig. 3. Corin is localized to human heart myocytes by immunostaining. Immunohistochemical staining of human heart tissues was performed using the affinity purified anticorin peptide A1 or A2 polyclonal antibodies as primary antibodies. (A) a longitudinal section of a representative heart tissue from a transplant recipient (Patient #7684) stained with AbA1 showing intense staining in the cardiac myocytes; (B) as (A) except the primary antibody was preadsorbed with the immunogenic peptide, A1, for 2 h; (C) the same tissue as (A) except stained with the weaker staining antibody, AbA2. Apparent staining at the poles of the nuclei are deposits of the brown lipochrome pigment, lipofuscin. (D) the same tissue as (A–C) processed in the absence of primary antibody; (E) a longitudinal section of normal myocardium from a heart which contained an acute infarct elsewhere (Patient #A4–99R) stained with AbA1 showing intense staining corresponding to the cross striations; (F) staining of the same heart tissue as (E) with AbA1 showing intense staining in cross section. Photomicrographs (A–E) were taken at an original magnification of 100 \times .

Anti-corin antibodies detect corin in heart lysates

We generated polyclonal antibodies to two different peptides derived from unique regions of the corin polypeptide sequence in order to investigate its expression and localization in the heart. The first was a unique region within the serine protease catalytic domain between the conserved Asp and Ser

amino-acid residues (AbA1) and the second was contained within the scavenger receptor domain (AbA2). Immunoblot analysis of corin protein expression in human heart protein lysates showed a major immunoreactive band of 125–135 kDa (Fig. 2), which was not present in lysates from the negative control HeLa cell line. This molecular mass is slightly lower than that reported (\approx 150 kDa) for recombinant V5/His6

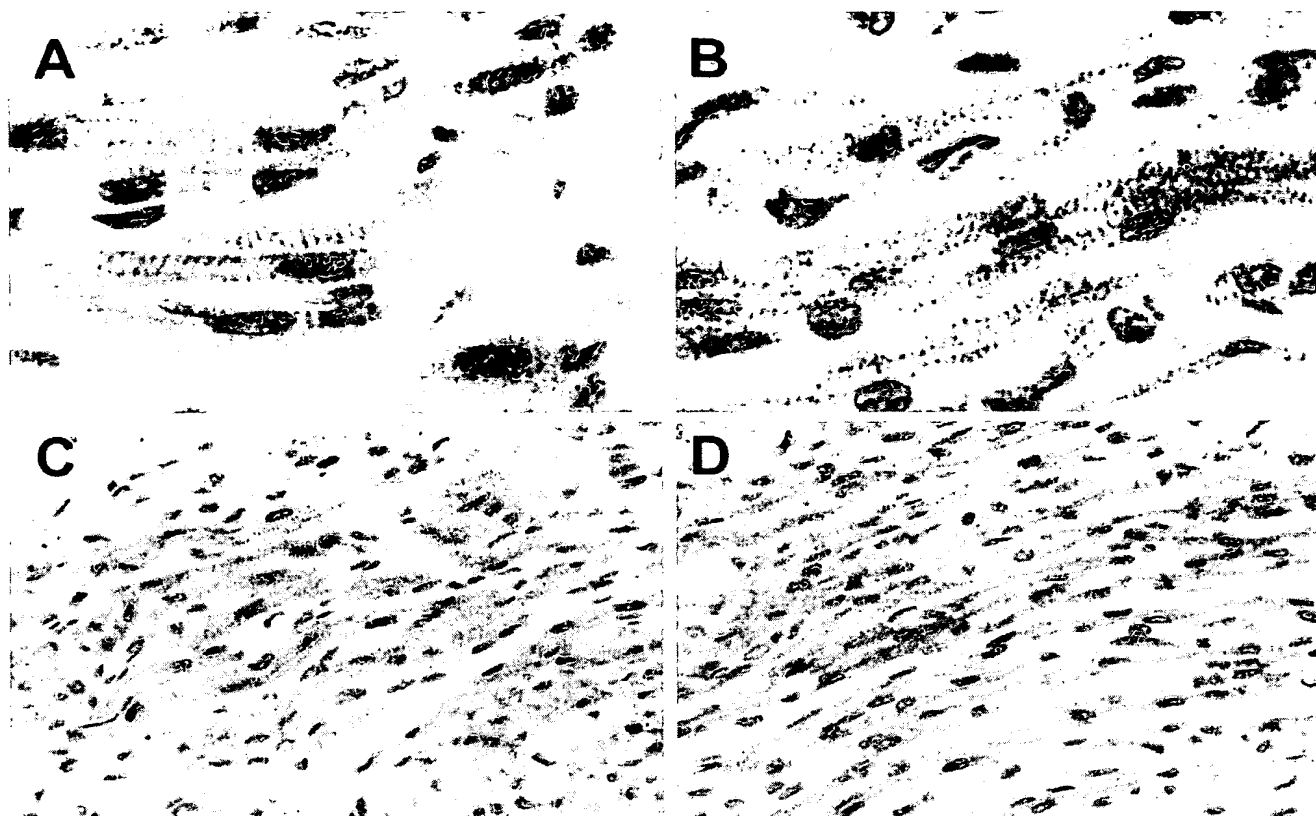


Fig. 4. Corin expression in neonate heart with TAPVR. Immunohistochemical staining of human neonate heart tissues was performed using the affinity purified anticorin peptide A1 polyclonal antibody as the primary antibody. (A) and (C) longitudinal sections of TAPVR heart tissue showing staining in the cardiac myocytes, corresponding to the cross striations; (B) and (D) longitudinal sections of a normal neonate heart showing a similar staining pattern in the cardiac myocytes. Photomicrographs (A) and (B) were taken at an original magnification of 100x and (C) and (D) were taken at an original magnification of 40x.

tagged corin expressed by human embryonic kidney 293 cells [20]. As the mature corin zymogen has a calculated mass of 116 kDa [19], it is likely that the mature corin polypeptide undergoes a post-translational processing event, possibly glycosylation. Consistent with this, there are 19 predicted N-linked glycosylation sites present in the extracellular domains of corin [19].

Corin is expressed by human heart myocytes

To investigate the localization of corin expression in human heart, immunohistochemical analyses were performed on human adult heart tissues. Corin was abundantly expressed in cardiac myocytes, with intense brown staining associated with cross striations seen in longitudinally sectioned myofibers (Fig. 3A). In some areas there was accentuation of the plasma membrane, consistent with an integral membrane localization of corin. This same pattern of staining was observed in sections taken from all areas of the myocardium. Control slides using the AbA1 polyclonal antibody in the presence of competing A1 peptide showed absence of this specific staining pattern (Fig. 3B). An identical, albeit weaker staining pattern was observed in experiments performed using the second corin-specific antibody (AbA2) (Fig. 3C). No staining was detected in the absence of antibody (Fig. 3D). Staining of a section of

viable myocardium from a heart containing an acute myocardial infarct showed a similar intense staining of the striations in cardiac myocytes (Fig. 3E) and a pinhead-like dot pattern when viewed in cross section (Fig. 3F). Necrotic heart tissue showed similar but much less intense staining (data not shown). Corin was not detected in sections of skeletal or smooth muscle (data not shown), suggesting that the function of corin is specifically related to cardiac muscle.

Corin protein expression in a patient with the congenital heart disease, TAPVR

The molecular mechanisms responsible for the developmental defect associated with the rare congenital heart disease TAPVR are not known. The location of the corin gene on human chromosome 4p12–13 [19] and the localization of the TAPVR locus to a 30 centimorgan interval on 4p13–q12 [26], suggested that corin may be a candidate for the TAPVR gene [19]. If corin plays a role in TAPVR, its expression may be lost or altered in TAPVR heart tissue. To explore this possibility, we examined corin protein expression in a TAPVR heart. The pattern of corin expression detected in this heart tissue (Fig. 4A,C) was similar to that observed in the adult heart and was identical to the pattern of corin staining in an age-matched neonate control heart (Fig. 4B,D). While this data is not consistent with a role

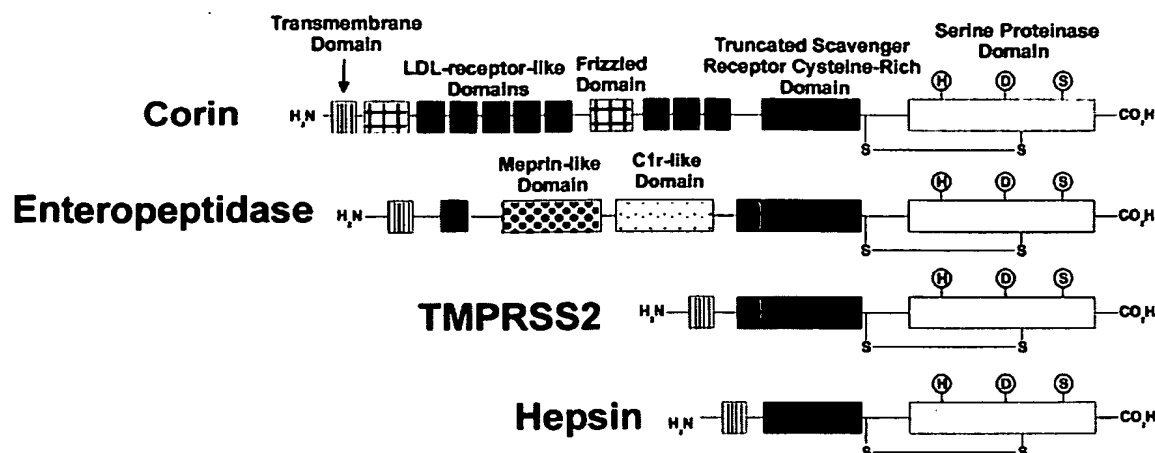


Fig. 5. Diagram showing domain structures of corin compared with other mosaic integral membrane proteins. The domains are as indicated. The catalytic serine protease residues are circled. The disulfide bond linking catalytic and pro-regions are marked.

for corin in TAPVR, it does not exclude the possibility that TAPVR is associated with more subtle alterations to the corin gene; for example point mutations, that would not be detected by this method.

Corin homology to other type II transmembrane proteases

As illustrated in Fig. 5, corin is a mosaic integral membrane protein possessing discrete domains. The intracellular, cytoplasmic domain contains two potential protein kinase C phosphorylation sites which may represent mechanisms for signal relay to or from the cell surface. Corin contains two frizzled domains. These domains function in other molecules as receptors for Wnt proteins, which are implicated in signal transduction during development [28]. Corin possesses eight LDL receptor domains which can mediate uptake of LDLs [29] and have also been shown to be involved in binding and internalization of protease/inhibitor complexes [30]. LDLs regulate the transport of cholesterol and play a major role in the development of heart disease. Corin possesses a scavenger receptor domain, which in other proteins, binds polyanionic molecules including modified lipoproteins, cell surface lipids and some sulfated polysaccharides [31]. The trypsin-like serine protease domain is located at the C-terminus.

Corin bears similarity to other known members of the integral membrane serine proteases as illustrated in Fig. 5. The corin serine protease domain is highly homologous to a multidomain integral-membrane serine protease found in the brush border of the intestine, enteropeptidase [32]. Enteropeptidase functions to activate digestive pancreatic enzymes released from the intestine. Activation of this cascade is critical, as illustrated by the life-threatening intestinal malabsorption that accompanies congenital deficiency of enteropeptidase [32]. Other proteases with homology to the corin serine protease domain are the integral-membrane serine proteases, TMPRSS2 and hepsin. Hepsin is a hepatic serine protease that has been demonstrated to activate Factor VII in the extrinsic blood coagulation pathway leading to thrombin formation, and has further been shown to be required for mammalian cell growth [33].

In summary, we have confirmed heart as a site of abundant corin mRNA expression and demonstrated for the first time the expression of corin as a 125–135 kDa protein in this tissue. In

addition, in heart we have localized corin protein to myocytes; the same cardiac cells expressing pro-ANP. These data support recently reported *in vitro* evidence that the corin proteolytic domain is the pro-ANP convertase [20] and thus, the proposal that corin has a role in regulating blood pressure. Possible additional functions of the serine protease domain and the functions of the other corin domains are not yet known. The putative phosphorylation sites in the cytoplasmic domain of corin may indicate that the intracellular domain of corin will be a target for phosphorylation and therefore may mediate signalling events from the cell surface. A better understanding of the role of corin in heart will provide insight into basic molecular mechanisms of cardiac function and could provide a rational target for both diagnostic and therapeutic applications.

ACKNOWLEDGEMENTS

This work was supported by grants from the Queensland Cancer Fund, Brisbane, Australia and the National Health and Medical Research Council of Australia. J. D. H. was supported by a John Earnshaw Scholarship from the Queensland Cancer Fund and by the Bancroft Scholarship, Queensland Institute of Medical Research.

REFERENCES

1. Rawlings, N.D. & Barrett, A.J. (1994) Families of serine peptidases. *Methods Enzymol.* **244**, 19–61.
2. Murphy, G. & Gavrilovic, J. (1999) Proteolysis and cell migration: creating a path? *Curr. Opin. Cell Biol.* **11**, 614–621.
3. LeMosy, E.K., Hong, C.C. & Hashimoto, C. (1999) Signal transduction by a protease cascade. *Trends Cell Biol.* **9**, 102–107.
4. Rifkin, D.B., Mazzieri, R., Munger, J.S., Noguera, I. & Sung, J. (1999) Proteolytic control of growth factor availability. *Acta Path. Microbiol. Immunol. Scand.* **107**, 80–85.
5. Dery, O. & Bunnett, N.W. (1999) Proteinase-activated receptors: a growing family of heptahelical receptors for thrombin, and trypsin. *Biochem. Soc. Trans.* **27**, 246–254.
6. Noel, A., Gilles, C., Bajou, K., Devy, L., Kebers, F., Lewalle, J.M., Maquoi, E., Munaut, C., Remacle, A. & d Foidart, J.M. (1997) Emerging roles for proteinases in cancer. *Invasion Metastasis* **17**, 221–239.
7. Ichinose, A. & Davie, E.W. (1994) The Blood Coagulation Factors: Their cDNAs, Genes, and Expression. In *Hemostasis and Thrombosis: Basic Principles and Clinical Practice* (Colman, R.W.,

- Hirsh, J., Marder V.J. & Salzman, E.W., eds), pp. 19–54. J.B. Lippincott Company, Philadelphia, PA, USA.
8. Francis, C.W. & Marder, V.J. (1994) Physiologic Regulation and Pathologic Disorders of Fibrinolysis. In *Hemostasis and Thrombosis: Basic Principles and Clinical Practice* (Colman, R.W., Hirsh, J., Marder V.J. & Salzman, E.W., eds), pp. 1076–1103. J.B. Lippincott Company, Philadelphia, PA, USA.
9. Arlaud, G.J. & Thielens, N.M. (1993) Human complement serine proteases C1r and C1s and their proenzymes. *Methods Enzymol.* **223**, 61–82.
10. Kitamoto, Y., Veile, R.A., Donis-Keller, H. & Sadler, J.E. (1995) Human complement serine proteases C1r and C1s and their proenzymes. *Biochemistry* **34**, 4562–4568.
11. Tsuji, A., Torres-Rosado, A., Arai, T., Le Beau, M.M., Lemons, R.S., Chou, S.H. & Kurachi, K. (1991) Hepsin, a cell membrane-associated protease. Characterization, tissue distribution, and gene localization. *J. Biol. Chem.* **266**, 16948–16953.
12. Paoloni-Giacobino, A., Chen, H., Peitsch, M.C., Rossier, C. & Antonarakis, S.E. (1997) Cloning of the TMPRSS2 gene, which encodes a novel serine protease with transmembrane, LDLRA, and SRCR domains and maps to 21q22.3. *Genomics* **44**, 309–320.
13. Schussheim, A.E. & Fuster, V. (1997) Thrombosis, antithrombotic agents, and the antithrombotic approach in cardiac disease. *Prog. Cardiovascular Diseases* **40**, 205–238.
14. Balcells, E., Meng, Q.C., Johnson, W.H. Jr, Oparil, S. & Dell'Italia, L.J. (1997) Angiotensin II formation from ACE and chymase in human and animal hearts: methods and species considerations. *Am. J. Physiol.* **273**, H1769–H1774.
15. Wolny, A., Clozel, J.P., Rein, J., Mory, P., Vogt, P., Turino, M., Kiowski, W. & Fischli, W. (1997) Functional and biochemical analysis of angiotensin ii-forming pathways in the human heart. *Circ. Res.* **80**, 219–227.
16. Bumpus, F.M. (1991) Angiotensin I and II. Some early observations made at the Cleveland Clinic Foundation and recent discoveries relative to angiotensin II Formation in human heart. *Hypertension* **18**, 122–125.
17. Kienast, J., Padro, T., Steins, M., Li, C.X., Schmid, K.W., Hammel, D., Scheld, H.H. & Van De Loo, J.C. (1998) Relation of urokinase-type plasminogen activator expression to presence and severity of atherosclerotic lesions in human coronary arteries. *Thromb. Haemost.* **79**, 579–586.
18. Labarrere, C.A., Pitts, D., Nelson, D.R. & Faulk, W.P. (1995) Vascular tissue plasminogen activator and the development of coronary artery disease in heart-transplant recipients. *N. Engl. J. Med.* **333**, 1111–1116.
19. Yan, W., Sheng, N., Seto, M., Morser, J. & Wu, Q. (1999) Corin, a mosaic transmembrane serine protease encoded by a novel cDNA from human heart. *J. Biol. Chem.* **274**, 14926–14935.
20. Yan, W., Wu, F., Morser, J. & Wu, Q. (2000) Corin, a transmembrane cardiac serine protease, acts as a pro-atrial natriuretic peptide-converting enzyme. *Proc. Natl Acad. Sci. USA* **97**, 8525–8529.
21. Sakanari, J.A., Staunton, C.E., Eakin, A.E., Craik, C.S. & McKerrow, J.H. (1989) Serine proteases from nematode and protozoan parasites: isolation of sequence homologs using generic molecular probes. *Proc. Natl Acad. Sci. USA* **86**, 4863–4867.
22. Elvin, C.M., Whan, V. & Riddles, P.W. (1993) A family of serine protease genes expressed in adult buffalo fly (*Haematobia irritans exigua*). *Mol. Gen. Genet.* **240**, 132–139.
23. Elvin, C.M., Vuocolo, T., Smith, W.J., Eisemann, C.H. & Riddles, P.W. (1994) An estimate of the number of serine protease genes expressed in sheep blowfly larvae (*Lucilia cuprina*). *Insect Mol. Biol.* **3**, 105–115.
24. Hooper, J.D., Nicol, D.L., Dickinson, J.L., Eyre, H.J., Scarman, A.L., Normyle, J.F., Stuttgen, M.A., Douglas, M., Loveland, K.A.L., Sutherland, G.R. & Antalis, T.M. (1999) Testisin, a new human serine proteinase expressed by premeiotic testicular germ cells and lost in testicular germ cell tumors. *Cancer Res.* **59**, 3199–3205.
25. Dickinson, J.L., Bates, E.J., Ferrante, A. & Antalis, T.M. (1995) Plasminogen activator inhibitor type 2 inhibits tumor necrosis factor alpha induced apoptosis. Evidence for an alternate biological function. *J. Biol. Chem.* **270**, 27894–27904.
26. Antalis, T.M. & Dickinson, J.L. (1992) Control of plasminogen activator inhibitor type 2 gene expression in the differentiation of monocytic cells. *Eur. J. Biochem.* **205**, 203–209.
27. Bleyl, S., Nelson, I., Odelbury, S.J., Ruttonberg, H.D., Otterud, B., Leppert, M. & Ward, K. (1995) A gene for familial total anomalous pulmonary venous return maps to chromosome 4p13-q12. *Am. J. Hum. Genetics* **56**, 408–415.
28. Cadigan, K.M. & Nusse, R. (1997) Wnt signaling: a common theme in animal development. *Genes Dev.* **11**, 3286–3305.
29. Bujo, H., Yamamoto, T., Hayashi, K., Hermann, M., Nimpf, J. & Schneider, W.J. (1995) Mutant oocyte low density lipoprotein receptor gene family member causes atherosclerosis and female sterility. *Proc. Natl Acad. Sci. USA* **92**, 9905–9909.
30. Kounnas, M.Z., Church, F.C., Argraves, W.S. & Strickland, D.K. (1996) Cellular internalization and degradation of antithrombin III-thrombin, heparin cofactor II-thrombin, and α 1-antitrypsin-trypsin complexes is mediated by the low density lipoprotein receptor-related protein. *J. Biol. Chem.* **271**, 6523–6529.
31. Resnick, D., Chatterton, J.E., Schwartz, K., Slayter, H. & Krieger, M. (1996) Structures of class A macrophage scavenger receptors. Electron microscopic study of flexible, multidomain, fibrous proteins and determination of the disulfide bond pattern of the scavenger receptor cysteine-rich domain. *J. Biol. Chem.* **271**, 26924–26930.
32. Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D.W. & Sadler, J.E. (1994) Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains. *Proc. Natl Acad. Sci. USA* **91**, 7588–7592.
33. Torres-Rosado, A., O'Shea, K.S., Tsuji, A., Chou, S.H. & Kurachi, K. (1993) Hepsin, a putative cell-surface serine protease, is required for mammalian cell growth. *Proc. Natl Acad. Sci. USA* **90**, 7181–7185.

Exhibit 11



US005645833A

United States Patent [19]**Dawson et al.**[11] **Patent Number:** **5,645,833**[45] **Date of Patent:** **Jul. 8, 1997**[54] **INHIBITOR RESISTANT SERINE
PROTEASES**[75] **Inventors:** **Keith Martyn Dawson; Richard
James Gilbert, both of Cowley, United
Kingdom**[73] **Assignee:** **British Biotech Pharmaceuticals
Limited, Oxford, United Kingdom**[21] **Appl. No.:** **379,621**[22] **PCT Filed:** **Aug. 3, 1993**[86] **PCT No.:** **PCT/GB93/01632**§ 371 Date: **Feb. 3, 1995**§ 102(e) Date: **Feb. 3, 1995**[87] **PCT Pub. No.:** **WO94/03614****PCT Pub. Date: Feb. 17, 1994**[30] **Foreign Application Priority Data****Aug. 4, 1992 [GB] United Kingdom 9216558**[51] **Int. Cl.⁶ A61K 38/48; C12N 9/68;
C12N 15/55; C12N 15/63**[52] **U.S. Cl. 424/94.64; 435/217; 435/252.3;
435/320.1; 435/325; 435/358; 435/365;
435/367; 435/369; 435/357; 435/352; 435/356;
536/23.2**[58] **Field of Search 424/94.64; 435/217,
435/172.3. 240.2, 252.3, 320.1; 536/23.2**[56] **References Cited****FOREIGN PATENT DOCUMENTS**

0381331	8/1990	European Pat. Off. .
WO9010649	9/1990	WIPO .
WO9109118	6/1991	WIPO .
WO9206203	4/1992	WIPO .

Primary Examiner—Dian C. Jacobson
Attorney, Agent, or Firm—Hale And Dorr[57] **ABSTRACT**

Serine proteases of the chymotrypsin superfamily are modified so that they exhibit resistance to serine protease inhibitors. If such modified serine proteases have fibrinolytic, thrombolytic, antithrombotic or prothrombotic properties, they are useful in the treatment of blood clotting diseases or conditions.

22 Claims, 18 Drawing Sheets

1 50

Complement Factor B WEHRKGTDYH KQPWQAKISV IRPSKGH..E SCMGAVVSEY FVLTAAHCF.

Complement C2 GVGNMSANAS DQERTPWHVT IKP.KSQ..E TCRGALISDQ WVLTAAHCF.

Medullasin IVGGRRARPH AWPFMVSLQL R...GG...H FCGATLIAPN FVMSAAHCV.

Myeloblastin MASLQM RGNPGS...H FCGGTLIHPS FVLTAAHCL.

Complement C1S IIGGSDADIK NFPWQVFF.. .D.NP.... WAGGALINEY WVLTAAHVV.

Complement C1R IIGGQKAKMG NFPWQVFT.. .NIHG.... RGGGALLGDR WILTAAHTL.

Factor X IVGGQECKDG ECPWQALLI. NEENEG.... FCGGTILSEF YILTAAHCL.

Factor IX VVGGEDAKPG QFPWQVVL.. NGKVDA.... FCGGSIVNEK WIVTAAHCV.

Factor VII IVGGKVCPKG ECPWQVLL. VNGAQ..... LCGGTILINTI WVVSAAHCF.

Protein C LIDGKMTRRG DSPWQVLL. DSKKKL.... ACGAVLIHPS WVLTAAHCM.

Thrombin IVEGSDAEIG MSPWQVMLFR KSPQEL.... LCGASLISDR WVLTAAHCLL

u-PA IIGGEFTTIE NOPWFAAIYR RH.RGGSVTY VCGGSLMSPC WVISATHCF.

t-PA IKGGLFADIA SHPWQAAIFA KHRRSPGERF LCGGILISSC WILSAAHCF.

Factor XII VVGGLVALRG AHPYIAALYW GHS..... FCAGSLIAPC WVLTAAHCL.

Apolipoprotein A IVGGCVAHPH SWPWQVSL.R .TRFGK...H FCGGTILISPE WVLTAAHCL.

Plasmin VVGCVVAHPH SWPWQVSL.R .TRFGM...H FCGGTILISPE WVLTAAHCL.

Hepsin IVGGRDTSIG RWPWQVSL.R .YD.GA...H LCGGSLLSGD WVLTAAHCF.

Elastase IIIa VVHGEDAVPY SWPWQVSL.Q .YEKSGSFYH TCGGSLIAPD WVVTAGHCI.

Elastase IIIB VVNGEDAVPY SWPWQVSL.Q .YEKSGSFYH TCGGSLIAPD WVVTAGHCI.

FIG.1A

1 50

Elastase IIa VVGGEARP N SWPQVSL.Q .YSSNGKWYH TCGGSLIANS WVLTAAHCI.

Elastase IIb MLGGEARP N SWPQVSL.Q .YSSNGQWYH TCGGSLIANS WVLTAAHCI.

Chymotrypsin B IVNGEDAVPG SWPQVSL.Q .DKTG...FH FCGGSLISED WVVTAAHCG.

Alpha Tryptase IVGGEAPRS KWPQVSL.R .VR.DRYMMH FCGGSLIHPQ WVLTAAHCL.

Beta Tryptase IVGGEAPRS KWPQVSL.R .VH.GPYMMH FCGGSLIHPQ WVLTAAHCV.

Factor XI IVGGTASVRG EWPQVTL.H .TT.SPTQRH LCGGSLIIGNQ WILTAAHCF.

Plasma Kallikrein IVGGTNSSWG EWPQVSL.Q .VK.LTAQRH LCGGSLIGHQ WVLTAAHCF.

Acrosin IVGKAAQHG AWPWMVSL.Q IFRYNSHRYH TCGGSLLSNR WVLTAAHCF.

Trypsin I IVGGYNCEEN SVPYQVSL.. ..NS.G..YH FCGGSLINEQ WVVSAGHCY.

Trypsin II IVGGYICEEN SVPYQVSL.. ..NS.G..YH FCGGSLISEQ WVVSAGHCY.

Trypsin III IVGGYTCEEN SLPYQVSL.. ..NS.G..SH FCGGSLISEQ WVVSAAHCY.

Tissue Kallikrein 2 IVGGWECEKH SQPWQVAV.. ..YSHG..WA HCGGVLVHPQ WVLTAAHCL.

PSA IVGGWECEKH SQPWQVLV.. ..ASRG..RA VCGGVLVHPQ WVLTAAHCI.

Tissue Kallikrein 1 IVGGWECEQH SQPQAAL.. ..YHFS..TF QCGGILVHRQ WVLTAAHCI.

Granzyme B IIGGHEAKPH SRPYMAYL.M IWDQKS..LK RCGGFLIQDD FVLTAAHCW.

T-cell Granzyme IIGGHEAKPH SRPYMAFV.Q FLQEK..RK RCGGILVRKD FVLTAAHCQ.

Cathepsin G IIGGRESRPH SRPYMAYL.Q IQSPAG..QS RCGGFLVRED FVLTAAHCW.

Complement Factor D ILGGREAEAH ARPYNASV.Q L...NG..AH LCGGVLVAEQ WVLSAAHCL.

Granzyme A IIGGNEVTPH SRPYMVLL.S L...DR..KT ICAGALIAKD WVLTAAHC..

Complement Factor I IVGGKRAQLG DLPWQVAIKD ASGIT..... .CGGIYIGGC WILTAAHCL.

FIG.1B

51 100

Complement Factor B ...TVDDKEH SI.KVSVGGE K....RDLEI EVVLFHPNYN INGKKEAGIP

Complement C2 ...R.DGN DH SLWRVNVGDP K SQWGKELLI EKAVISPGFD VFAKKNQGIL

MedullasinANVNV RAVRVVLGAH NLSRREPTRQ VFAVQRIFEN GYDPVNLL..

MyeloblastinRDIPO RLVNVVLGAH NVRTQETQQ HFSVAQVFLN NYDAENKL..

Complement C1SEGN REPTMYVGST SVQTSRLAKS KMLTPEHVFI HPGWKLLLEVP

Complement C1R YPKEHEAQSN ASLDVFLGHT NVEE..LMKL GNHPIRRVSU HPDYR....Q

Factor XYQAK.. .RFKVRVGDR NTEQEEGG.E AVHEVEVVVK HNRF.....

Factor IXETGV.. .KITVVAGEH NIEETEHT.E QKRN VIRIIP HHNVNA.....

Factor VIIDKIKNW RNLI AVLGEH DLSEHDG.D E QSRRAQVII PSTYVP....

Protein CDESK.. .KLLVRLGEY DLRRWEKW.E LDLDIKEVVF HPNY.....

Thrombin YPPWDKNFTE NDLLVRIGKH SRTRYERNIE KISMLEKIYI HPRYNW....

u-PA .IDYPKKE.. .DYIVYLGRS RLNSNTQGEK KF..... .EVENLILH

t-PA .QERFPPH.. .HLTIVILGRT YRVVPGEETQ KF..... .EVEKYIVH

Factor XII .QDRPAPE.. .DLTVVLGOE RRNHSCEPCQ TL..... .AVRSYRLH

Apolipoprotein A .K..KSSRP. SSYKVVILGAH QEV...NLES HV.....QE. .IEVSRLFL

Plasmin .E..KSPRP. SSYKVVILGAH QEV...NLEP HV.....QE. .IEVSRLFL

Hepsin .P..ERNRVL SRWRVFAGAV AQASPHGLQL GV.....QA. .WYHGGYL

Elastase IIIa .S..RD.... LTYQVVLGEY NLAVKEGPEQ VI.....PI. .NSEELFVH

Elastase IIIB .S..SS.... RTYQVVLGEY DRVKEGPEQ VI.....PI. .NSGDLFVH

Elastase IIA .S..SS.... RTYRVGLGRH NLYVAESGSL AV.....SV.SKIVVH

FIG. 1C

51 100

Elastase IIB .S..SS.... RIYRVMGQH NLYVAESGL AV.....SV.SKIVVH

Chymotrypsin B .V..RT.... SDV.VVAGEF DQGSDEENIQ VL.....KI.AKVFKN

Alpha Tryptase .G..PDVKDL ATRLVN.SGT HLYYQDQLLP VS.....RI. ..MVHPQFYI

Beta Tryptase .G..PDVKDL AALRVOLREQ HLYYQDQLLP VS.....RI. ..IVHPQFYT

Factor XI .YGVESPKIL RVYSGILNQS EIKEDTSFFG VQ.....EI. ..IIHDQYKM

Plasma Kallikrein .DGLPLQDVW RIYSGILNLS DITKDTFFSQ IK.....EI. ..IIHQNYKV

Acrosin .VGKNNVHD. ..WRLVFGAK EITYGNKPV KA.....PLQ ERYVEKIIH

Trypsin IKSRI. ...QVRLGEH NIEVLEGNEQ F.INAAKIIR HPQYDRKTIN

Trypsin IIKSRI. ...QVRLGEH NIEVLEGNEQ F.INAAKIIR HPKYNSTRILD

Trypsin IIIKTRI. ...QVRLGEH NIKVLEGNEQ F.INAAKIIR HPKYNRDTLD

Tissue Kallikrein 2KKNS. ...QVWLGRH NLFEPEDTGO R.VPVSHSFP HPLYNMSLLK

PSARNKS. ...VILLGRH SLFHPEDTGO V.FQVSHSFP HPLYDMSLLK

Tissue Kallikrein 1SDNY. ...QLWLGRH NLFDDENTAQ F.VHVSESFP HPGFNMSLLE

Granzyme BGSSINVTILGAH NIKEQEPTQQ F.IPVKRPPI HPAYNPKNFS

T-cell GranzymeGSSINVTILGAH NIKEQEPTQQ F.IPVKRPPI HPAYNPKNFS

Cathepsin GGSNINVTILGAH NIQRRENTQQ H.ITARRAIR HPQYNQRTIQ

Complement Factor DEDAAD GKQVVLIGAT HLPQPEPXXX ITIEVLRAVP HPDSQPDITD

Granzyme ANLN KRSQVILGAH SITREPTKO IML.VKKEFP YPCYDPATRE

Complement Factor IRASKT HRYQIWTTVV DWIHPDLKRI VIEYVDRIIF HENYNA....

FIG.1D

101

150

Complement Factor B EFY..... DYDVALIKL.KNKLY QGIRPICLP CTEGTRALR
 Complement C2 EFY..... GDDIALKL.AQVKM STHARPICLP CTMEANLALR
 MedullasinNDIVILQL.NGSATI NANVQAQLP AQGR....RL
 MyeloblastinNDILLIQL.SSPANL SASVTSVQLP QQDQ....PV
 Complement C1S E....GRTNF DNDIALVRL.KDPVKM GPTVSPICLP GTSSDYNLMD
 Complement C1R D....ESYNF EGDIALLEL.ENSVTL GPNLLPICLP DNDTFYDL..
 Factor XTKETY DFDIAVLRL.KTPITF RMNVAPACLP ERDWAESTL.
 Factor IXAINKY NHDIALLEL.DEPLVL NSVTPICIA DKEYTN.IF.
 Factor VIIG..TT NHDIALLRL.HQPWVL TDHVVPLCLP ERTFSERTL.
 Protein CSKSTT DNDIALLHL.AQPATL SQTIVPICLP DSGLAEREIN
 ThrombinRENL DRDIALMKL.KKPVAF SDYIHPVCLP DRETAASLLQ
 u-PA KDYSADTLAH HNDIALLKIR SK.EGRCAQP SRTIQTICLP SMY...NDPQF
 t-PA KEFDDDT..Y DNDIALLLQLK SD.SSRCAQE SSVVRTVCLP P....ADLQL
 Factor XII EAFS..PVS Y QHDLALLRLQ EDADGSCALL SPYVQPVCLP SGA...ARP..
 Apolipoprotein AEPT QADIALKL.SRPAV.I TDKVMPACLP SPD..YMT.
 PlasminEPT RKDIALKL.SSPAV.I TDKVIPACLP SPN..YVVA.
 Hepsin PFRDPNSEN SNDIALVHL.SSPLP.L TEYIQPVCLP AAG..QALV.
 Elastase IIIa PLWNRSCVAC GNDIALIKL.SRSAQ.L GDAVQLASLP PAG..DILP.
 Elastase IIIb PLWNRSCVAC GNDIALIKL.SRSAQ.L GDAVQLASLP PAG..DILP.
 Elastase IIa KDWSNQISK GNDIALKL.ANPVS.L TDKIQLACLP PAG..TILP.

FIG.1E

101

150

Elastase IIB KDWNSQVSK GNDIALKL. ...ANPVS.L TDKIQLACLP PAG..TILP.
 Chymotrypsin B PKF..SILTV NNDITLLKL. ...ATPAR.F SQTSAVCLP SAD..DDFP.
 Alpha TryptaseIQT GADIALLEL. ...EEPVN.I SSRVHTVMPL PAS..ETFP.
 Beta TryptaseAQI GADIALLEL. ...EEPVK.V SSHVHTVTLP PAS..ETFP.
 Factor XIAES GYDIALKL. ...ETVN.Y TDSQRPICLP SKG..DRNV.
 Plasma KallikreinSEG NHDIALIKL. ...QAPIN.Y TEFQKPICLP SKG..DTST.
 Acrosin EKYS..ATE GNDIALVEI. ...TPPIS.C GRFIGPGCLP HFK..AGLP.
 Trypsin I N..... ..DIMLIK. ...SSRA.VI NARVSTISLP TAP..PAT..
 Trypsin II N..... ..DILLIK. ...SSPA.VI NSRVSAISLP TAP..PAA..
 Trypsin III N..... ..DIMLIK. ...SSPA.VI NARVSTISLP TAP..PAA..
 Tissue Kallikrein 2 HQSLRPDEDS SHDLMLRL. ...SEPAK.I TDVVKVGLP TQE..PAL..
 PSA NRFLRPGDDS SHDLMLRL. ...SEPAE.L TDAVKVMDLP TQE..PAL..
 Tissue Kallikrein 1 NHTRQADEY SHDLMLRL. ...TEPADTI TDAVKVVELP TQE..PEV..
 Granzyme B N..... ..DIMLIQL. ...ERKAK.R TRAVQPLRLP SNK..AQVK.
 T-cell Granzyme N..... ..DIMLIQL. ...ERKAK.W TTAVRPLRLP SSK..AQVK.
 Cathepsin G N..... ..DIMLIQL. ...SRRVR.R NRVNPFVALP RAQ..EGLR.
 Complement Factor D H..... ..DLLLLQL. ...SEKAT.L GPAVRPLPWQ RVD..RDVA.
 Granzyme A G..... ..DLKLIQL. ...TEKAK.I NKVVTILHLP KKG..DDVK.
 Complement Factor IGTY QNDIALIEMK KDGKCKOCELPRSIP ACVPWSPYLF

FIG.1F

151 200

Complement Factor B LPPTTTCQQ KEELLPAQDI KALFVSEEEK KLTRKEVYIK NGDKKGSC.E

Complement C2 RPOGSTCRDH ENELLNKQSV PAHFVALNGS KL...NINLK MGVEWTS CAE

Medullasin GNGVQCLAMG WGLL...GRNRGIASVL QELNVTV...VT.....

Myeloblastin PHGTQCLAMG WGRV.....GAHDPPAQVL QELNVTV...VT.....

Complement C1S GDL..GLISG WGRTEK....RDRAVRL KAARLPV...APLRKCKE

Complement C1R GLM..GYVSG FGVMEE....KI.AHDL RFVRLPV...ANPOACEN

Factor X MTQKTGIVSG FGRTHE.KGR QS....TRL KMLEVPY...VDRNSCKL

Factor IX LKFGSGYVSG WGRVFH.KGR SA....LVL QYLRVPL...VDRATCLR

Factor VII AFVRFSLVSG WGQLLD.RGA TA....LEL MVLNVPR...LMTQDCLQ

Protein C QAGQETLVTG WGYHSS.REK EAKRNRTFVL NFIKIPV...VPHNECSE

Thrombin AGYK.GRVTG WGNLKETWTA NVGKGQPSVL QVVNLPI...VERPVCKD

u-PA G..TSCEITG FGKENS....TDYLYPEQ.L KMTVVKL...ISHRECQQ

t-PA PDWTECELSG YGKHEA....LSPFYSER.L KEAHVRL...YPSSRCTS

Factor XII SETTLCQVAG WGHQFE....GAEEYASF.L QEAQVPF...LSLERC SA

Apolipoprotein A ARTE.CYITG WGETQG....TFG..TG.LL KEAQLLV...IENEVCNH

Plasmin DRTE.CFITG WGETQG....TFG..AG.LL KEAQLPV...IENKVCNR

Hepsin DGKI.CTVTG WGNTQ.....YYGQQAG.VL QEARVPI...ISNDVCNG

Elastase IIIa NKTP.CYITG WGRLYT....NGP.LPD.KL QQARLPV...VDYKHCSR

Elastase IIIb NETP.CYITG WGRLYT....NGP.LPD.KL QEALLPV...VDYEHCSR

Elastase IIa NNYP.CYVTG WGRLOT....NGA.VPD.VL QQGRLLV...VDYATCSS

FIG.1G

151	200
Elastase Iib	NNYP.CYVTG WGRLOT.... NGA.LPD.DL KQGRLLV... ..VDYATCSS
Chymotrypsin B	AGTL.CATTG WGKTKY.... NANKTPD.KL QQAALPL... ..LSNAECKK
Alpha Tryptase	PGMP.CWVTG WGDVDN.... DEPLPPFPPL KQVKVPI... ..MENHICDA
Beta Tryptase	PGMP.CWVTG WGDVDN.... DERLPPFPPL KQVKVPI... ..MENHICDA
Factor XI	IYTD.CWVTG WGYRKL.... RDKIQN..TL QKAKIPL... ..VTNEECQK
Plasma Kallikrein	IYTN.CWVTG WGFSKE.... KGEIQN..IL QKVNIP... ..VTNEECQK
Acrosin	RGSQSCWVAG WGYIEE.... KAP.RPSSIL MEARVDL... ..IDL DLCNS
Trypsin I	.GTK.CLISG WGNTAS.... SGADYPD.EL QCLDAPV... ..LSQAKCEA
Trypsin II	.GTE.SLISG WGNTLS.... SGADYPD.EL QCLDAPV... ..LSQAECEA
Trypsin III	.GTE.CLISG WGNTLS.... FGADYPD.EL KCLDAPV... ..LREAECKA
Tissue Kallikrein 2	.GTT.CYASG WGSIEP.... EEFLRPR.SL QCVSLHL... ..LSNDMCMAR
PSA	.GTT.CYASG WGSIEP.... EEFLTPK.KL QCVDLHV... ..ISNDVCAQ
Tissue Kallikrein 1	.GST.CLASG WGSIEP.... ENFSFPD.DL QCVDLKI... ..LPNDECEK
Granzyme B	PGQT.CSVAG WQTAP.... LG.KHSH.TL QEVKMTV... ..QEDRKCES
T-cell Granzyme	PGQL.CSVAG WG.YVS.... MS.TLAT.TL QEVLLTV... ..QKDCQCER
Cathepsin G	PGTL.CTVAG WGR.VS.... MR.RGTD.TL REVQLRV... ..QRDRQCLR
Complement Factor D	PGTL.CDVAG WGIVNH.... AG.RRPD.SL QHVLLPV... ..LDRATCRL
Granzyme A	PGTM.CQVAG WGRTHN.... SA.SWSD.TL REVNITI... ..IDRKVCND
Complement Factor I	QPNDTCIVSG WGREKDNERV FSLQWGEVKL ISNCSKF... ..YGNRFYEK

FIG. 1H

201

250

Complement Factor B RDAQYAPGYD KVKDISEVVT PRFLCTGGVS PYADPNTCRG DSGGPLIVHK
Complement C2 VVSQEKTMFP NLTDVREVVT DQFLCSGTQ. . . EDESPCKG ESGGAVFLER
Medullasin SL CRRSNVCTLV RGRQAGVCFG DSGSPLVCNG
Myeloblastin FF CRPHNICTFV PRRKAGICFG DSGGPLICDG
Complement C1s VKVEKPTADA EAYVFTPNMI CAG. GEK G. MDSCKG DSGGAFVQD
Complement C1r WLRGKNRMD. VFSQNMF CAGH. PSL K. QDACQG DSGGVFAVRD
Factor X SSSFI. ITQNMF CAGY. DTK Q. EDACQG DSGGPHV. . . T
Factor IX STKFT. IYNNMF CAGF. HEG G. RDSCQG DSGGPHV. . . T
Factor VII QSRKVG DSPNITEYMF CAGY. SDG S. KDSCKG DSGGPHA. . . T
Protein C VMSNM. VSENML CAGI. LGD R. QDACEG DSGGPMV. . . A
Thrombin STRI. RITDNMF CAGYKPDGK R. GDACEG DSGGPFVMKS
u-PA PHYYS. EVTTKML CAADPQWKT. DSCQG DSGGPLVCSL
t-PA QHLLNR. TVTDNML CAGDTRSGGP QANLHDACQG DSGGPLVCLN
Factor XII PDVHGS. SILPGML CAGFLEGGT. DACQG DSGGPLVCED
Apolipoprotein A YKY. I CAEHLARGT. DSCQG DSGGPLVCFE
Plasmin YEFLNG. RVQSTEL CAGHLAGGT. DSCQG DSGGPLVCFE
Hepsin ADFYGN. QIKPKMF CAGYPEGGI. DACQG DSGGPFVCE
Elastase IIIa WNWGS. TVKKTIV CAG.GY.IR. SGCNG DSGGPLNCPT
Elastase IIIB WNWGS. SVKKTIV CAG.GD.IR. SGCNG DSGGPLNCPT
Elastase IIA SAWGS. SVKTSMI CAG.GD.GVI. SSCNG DSGGPLNCQA

FIG.1I

201 250

Elastase IIb SGWGS.... ...TVKTNMI CAG.GDGI.CTCNG DSGGPLNCQA
Chymotrypsin B S..WGR.... ...RITDVM CAG.ASGV..SSCMG DSGGPLVCQ.
Alpha Tryptase KYHLGAYTGD DVRIIRDML CAG..NSQR.DSCKG DSGGPLVCKV
Beta Tryptase KYHLGAYTGD DVRIVRDML CAG..NTRR.DSCQ DSGGPLVCKV
Factor XI RYR..... .GHKITHMI CAGYREGGK.DAACKG DSGGPLSCKH
Plasma Kallikrein RYQ..... .DYKITQRMV CAGYKEGGK.DAACKG DSGGPLVCKH
Acrosin TQWYNG.... ...RVQPTNV CAGYPVGKI.DTCQ DSGGPLMCKD
Trypsin IS..... YPGKITSNMF CVGFLEGGK.DSCQ DSGGPVVCNG
Trypsin IIS..... YPGKITNNMF CVGFLEGGK.DSCQ DSGGPVVSNG
Trypsin IIIS..... CPGKITNSMF CVGFLEGGK.DSWKR DSGGPVVCNG
Tissue Kallikrein 2A..... YSEKVTEFML CAGLWTGGK.DTCGG DSGGPLVCNG
PSAV..... HPQKVTKFML CAGRWTGGK.STCSG DSGGPLVCNG
Tissue Kallikrein 1A..... HVQKVTDFML CVGHLEGGK.DTCVG DSGGPLMCDG
Granzyme B DLRHY..... YDSTIEL... CVGDPEIKK.TSFKG DSGGPLVCNK
T-cell Granzyme LFHGN..... YSRATEI... CVGDPKKTQ.TGFKG DSGGPLVCKD
Cathepsin G IF.GS..... YDPRRQI... CVGDRRERK.AAFKG DSGGPLLCNN
Complement Factor D YD..... VLRML CAESNR..R.DSCKG DSGGPLVCGG
Granzyme A RNHYN..... FNPVIGMNV CAGSLRGR.DSCNG DSGSPLCEG
Complement Factor IEME CAGTYDGI.DAACKG DSGGPLVCMD

FIG. 1J

251 300

Complement Factor B RS....RFIQ VGVISGWVD VC...KNQKR QKQVP....A HARDFHINLF

Complement C2 RF....RFFQ VGLVSWGLYN PCLGSADKNS RKRAPRSKVP PPRDFHINLF

MedullasinLI HGIASFVR.G GCASGLYPDA FAPVA.....

MyeloblastinII QGIDSFVI.W GCATRLFPDF FTRVA.....

Complement C1S PN.DKTKFYA AGLVSWGP.. QCG.T..YGL YTRVK.....

Complement C1R PN.TD.RWVA TGIVSWGII.. GCSRG..YGF YTKVL.....

Factor X RF.KDTYFV. TGIVSWGE.. GCARKGKYGI YTKVT.....

Factor IX EV.EGTSFL. TGIISWGE.. ECAMKGKYGI YTKVS.....

Factor VII HY.RGTWYL. TGIVSWGQ.. GCATVGHFV YTRVS.....

Protein C SF.HGTWFL. VGLVSWGE.. GCGLLHNYGV YTKVS.....

Thrombin PF.NNRWYQ. MGIVSWGE.. GCDRDGKYGF YTHVF.....

u-PA Q.G...RMTL TGIVSWGR.. GCALKDKPGV YTRVS.....

t-PA D.G...RMTL VGIISWGL.. GCGQKDVPGV YTKVT.....

Factor XII Q.AAERRLT QGIISWGS.. GCGDRNKPGV YTDVA.....

Apolipoprotein AKDKYIL QGVTSWG..L GCARPKNKPGV YARVS.....

PlasminKDKYIL QGVTSWG..L GCARPKNKPGV YVRVS.....

Hepsin SISRTPRWRL CGIVSWG..T GCALAQKPGV YTKVS.....

Elastase IIIa E...DGGWQV HGVTSFVSFAF GCNFIWKPTV FTRVS.....

Elastase IIIb E...DGGWQV HGVTSFVSFAF GCNTRRKPTV FTRVS.....

Elastase IIa S...DGRWQV HGIVSFGSRL GCNYYHKPSV FTRVS.....

FIG.1K

251

300

Elastase Iib S...DGRWEV HGIGSLTSVL GCNYYKKPSI FTRVS.....
Chymotrypsin B K...DGAWTL VGIVSWGSDT CST...SSPGV YARVT.....
Alpha TryptaseNGTWLQ AGVVSWE... GCAQPNRPGI YTRVT.....
Beta TryptaseNGTWLQ AGVVSWE... GCAQPNRPGI YTRVT.....
Factor XINEVWHL VGITSWGE... GCAQERPGV YTNVV.....
Plasma KallikreinNGMWRL VGITSWGE... GCARREQPGV YTKVA.....
Acrosin S..KESAYVV VGITSWG..V GCALAKRPGI YTATW.....
Trypsin IQL QGVVSWG DG. .CAQKNKPGV YTKV.....Y
Trypsin IIEL QGIVSWG YG. .CAQKNRPGV YTKV.....Y
Trypsin IIIQL QGVVSWG HG. .CAWKNRPGV YTKV.....Y
Tissue Kallikrein 2VL QGITSWG PE. PCALPEKPAV YTKV.....V
PSAVL QGITSWG SE. PCALPERPSL YTKV.....V
Tissue Kallikrein 1VL QGVTSWG YV. PCGTPNKPSV AVR.....L
Granzyme BVA QGIVSYGRNN GMP....PRA CTKVS.....
T-cell GranzymeVA QGILSYGNKK GTP....PGV YIKVS.....
Cathepsin GVA HGIVSYGKSS GVP....PEV FTRVS.....
Complement Factor DVL EGVVTSG.SR VCGNRKKPGI YTRVA.....
Granzyme AVF RGVTSPFLEN KCGDPRGPGV YILLS.....K
Complement Factor I ANNVTVW.. .GVVSWGE.. NCGKPEFPV YTKVA.....

FIG. 1L

301 350

Complement Factor B QVLPWLKEKL QEDLGL..

Complement C2 RMQPWLRLQHL GDVNLFLPL

Medullasin QFVNWIDSII QRSEDNPCPH PRODPASRT H.

Myeloblastin LYVDWIRSTL RRVEAKGRP.

Complement C1S NYVDWIMKTM QENSTPRED.

Complement C1R NYVDWIKKEM EEED.....

Factor X AFLKWIDRSM KTRGLPKAKS HAPEVITSSP LK.

Factor IX RYVNWIKET KLT.....

Factor VII QYIEWLQKLM RSEPRPGVLL RAPFP

Protein C RYLDWIHGHI RKEAPOKSW AP.....

Thrombin RLKKWIKQVI DQFGE.....

u-PA HFELWIRSH KEENGLAL..

t-PA NYLDWIRDNM RP.....

Factor XII YYLAWIREHT VS.....

Apolipoprotein A RFVTWIEGMM RNN.....

Plasmin RFVTWIEGVM RNN.....

Hepsin DFREWIFQAI KTHSEASGMV TQL.....

Elastase IIIa AFIDWIEETI ASH.....

Elastase IIIB AFIDWIEETI ASH.....

Elastase IIA NYIDWINSVI ANN.....

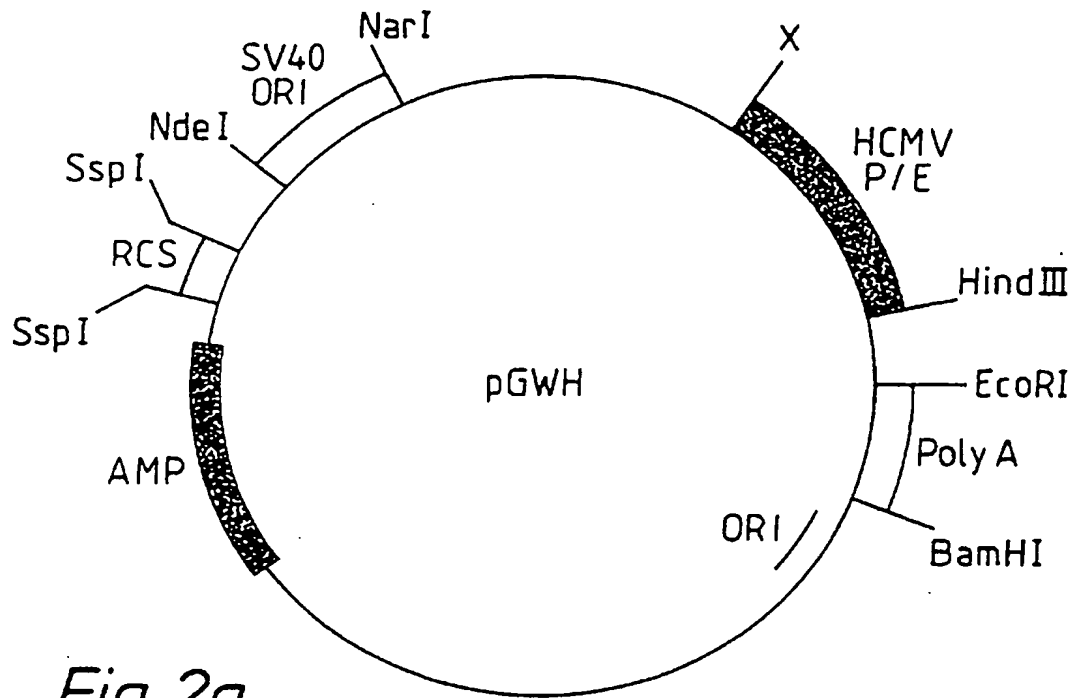
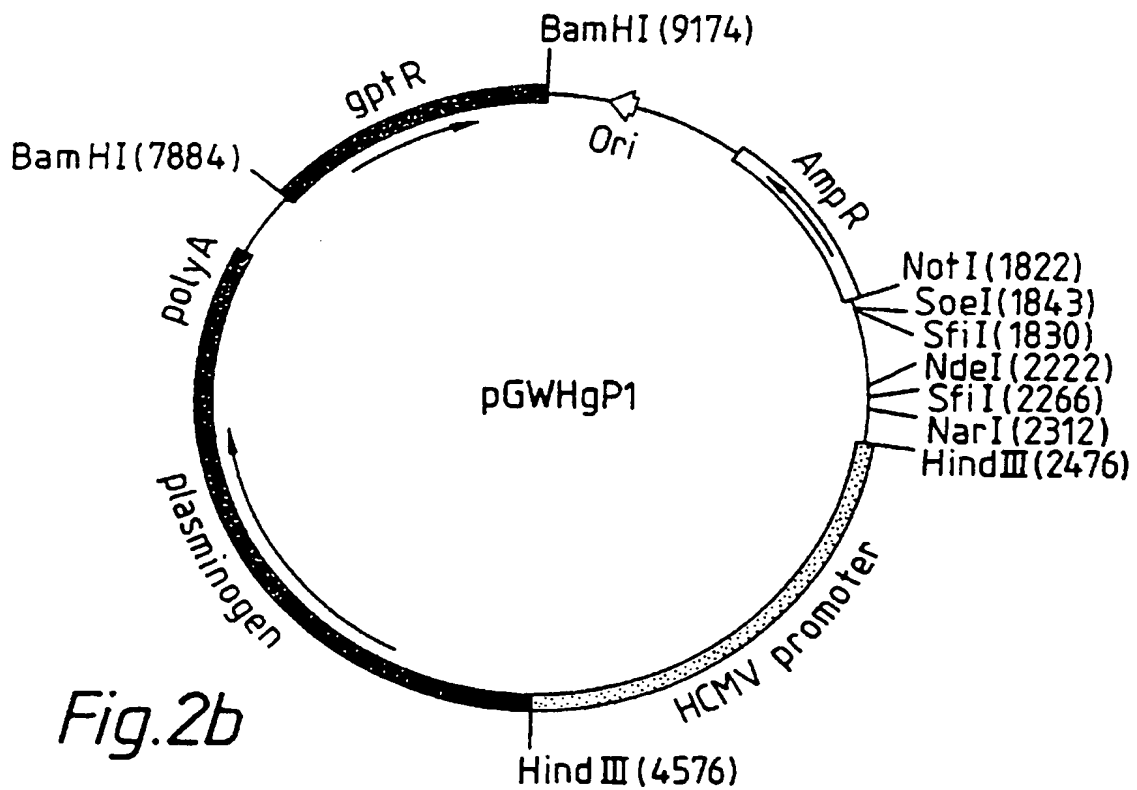
Elastase IIB NYNDWINSVI ANN.....

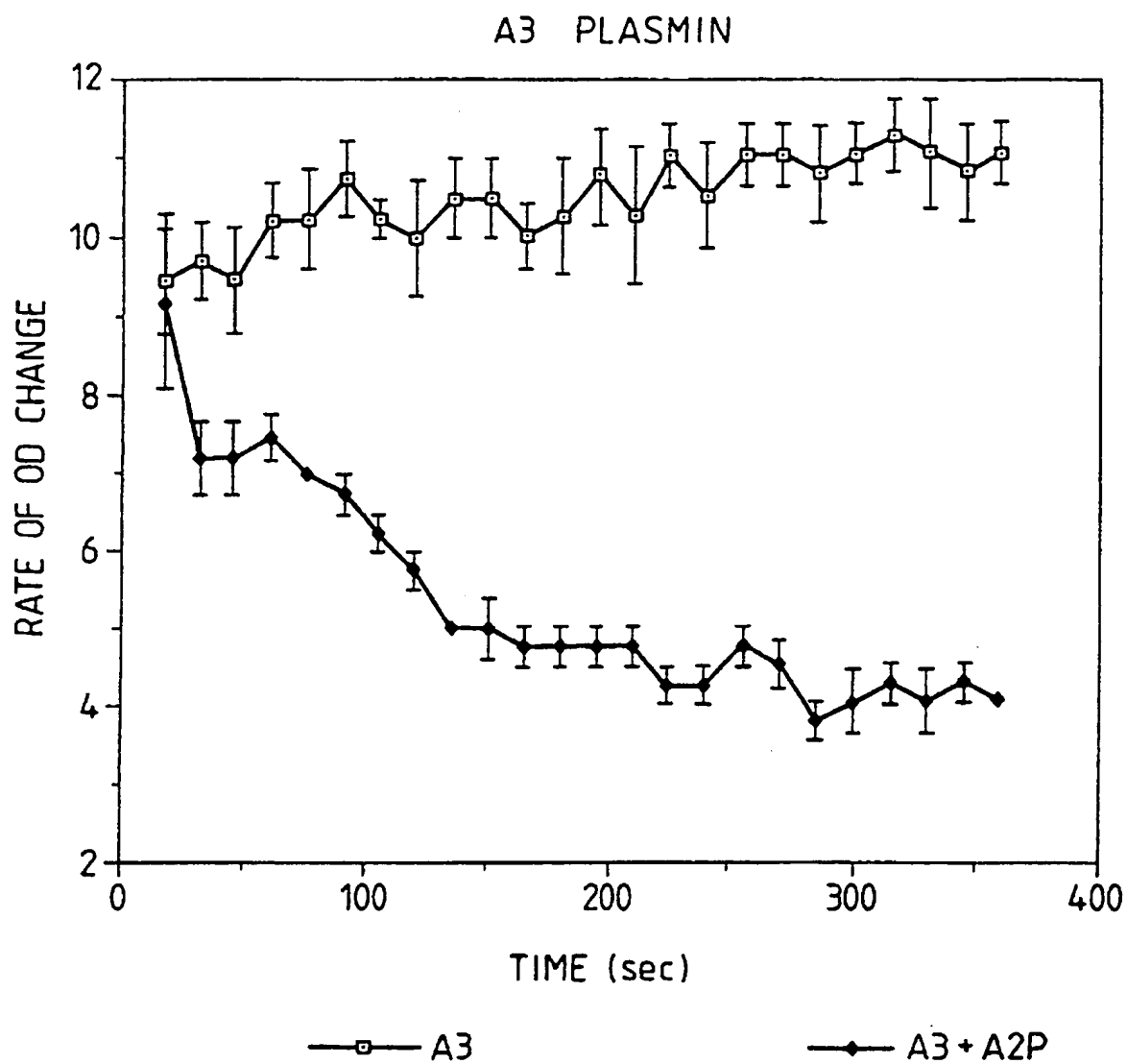
Chymotrypsin B KLIPWVQKIL AAN.....

FIG. 1M

301		350
Alpha Tryptase	YYLDWIHHYV PKKP.....
Beta Tryptase	YYLDWIHHYV PKKP.....
Factor XI	EYVDWILEKT QAV.....
Plasma Kallikrein	EYMDWILEKT QSSDGKAQMQ SPA.....
Acrosin	PYLNIWIASKI GSNALRMIOQ ATPPPPTTRP PPIRPPFSHP ISAHLPWFQ	
Trypsin I	NYVKWIKNTI AANS.....
Trypsin II	NYVDWIKDTI AANS.....
Trypsin III	NYVDWIKDTI AANS.....
Tissue Kallikrein 2	HYRKWIKDTI AANP.....
PSA	HYRKWIKDTI VANP.....
Tissue Kallikrein 1	SYVKWIEDTI AENS.....
Granzyme B	SFVHWIKKTM KRY.....
T-cell Granzyme	HFLPWIKRTM KRL.....
Cathepsin G	SFLPWIRTMT RSFKLLDQME TPL.....
Complement Factor D	TYAAWIDHVL
Granzyme A	KHLNWIIMTI KGAV.....
Complement Factor I	NYFDWISYHV GRPFISQYNV
	351	400
Acrosin	PPRPRLPPRP PAAQPPPPPS PPPPPPPAS PLPPPPPPPP PTPSSTTKLP	
	401	442
Acrosin	QGSLFAKRLQ QLIEVLKGKT YSDGKNHYDM ETTPELPTS TS	

FIG. 1N

*Fig. 2a**Fig. 2b*

*Fig. 3*

1	50
Antiplasmin	MALLWGLLVL SWSCLGGPCS VFSPVSAMEP LGRQLTSGPN QEQVSPLTLL
51	100
Antiplasmin	KLGNQEPGGQ TALKSPPGVC SRDPTPEQTH RLARAMMAFT ADLFSLVAQT
Ovalbumin	-----G SIGAASMEFC FDFVKELKVH
101	150
Antiplasmin	STCPNLILSP LSVALALSHL ALGAQNHTLQ RLQQVLHAGS GP-----
Ovalbumin	HANENIFYCP IAIMSALAMV YLGAKDSTRT QINKVVRFDK LPGFGDSIEA
151	200
Antiplasmin	-----CLPHLLSRLC QDLGPGAFL AARMYLQKGF PIKEDFLEQS
Ovalbumin	QCGTSVNVHS SLRDIILNQIT KPNQVYSFSL ASRLYAEERY PILPEYLQCV
201	250
Antiplasmin	EQLF--GAKP VSLTGKQEDD LANINQWKE ATEGKIOEFL S--GLPEDTV
Ovalbumin	KELYRGGLEP INFQTAADQA RELINSWVES QTNGIIRNVL QPSSVDSQTA
251	300
Antiplasmin	LLLLNAIHFO GFWRNKFDPS LTQRDSFILD EQFTVPVEMM -QARTYPLRW
Ovalbumin	MVLVNAIVFK GLWEKAFKOE DTQAMPFRVT EQESKPVQMM YOIGLFRVAS

FIG. 4A

301	350
Antiplasmin	FLLEQPEIQV AHFPFKNMS FVLVPTHFE WNVSQLANL SWOTLHPPLV
Ovalbumin	MASEKMKILE LPF-ASGTMS MLVLLPDEVS -GLEQLESII NFEKLTWTS
351	400
Antiplasmin	WE-----RPTK VRLPKLYLKH QMDLVATLSQ LGLQELF-QA PDLRGIS-EQ
Ovalbumin	SNVMEERKIK VYLPRMKMEE KYNLTSVLMA MGITDVFSS ANLSGISSAE
401	450
BBTI Loop	P CKARII
Antiplasmin	SLVSGVQHQ STLESEGV EAAATSIAM SRMSLSS-FS VNRPFLFFIF
Ovalbumin	SLKISQAVHA AHAEINEAGR EVGSAEAGV DAASVSEEF ADHPFLFCIK
451	500
Antiplasmin	EDTIGLPLFV GSVRNPNSA PRELKEQDS PGNKDFLQSL KGFPRGDKLF
Ovalbumin	HIATNAVLFF GRCVSP
501	521
Antiplasmin	GPLDKLVPPM EEDYPQFGSP K

FIG. 4B

INHIBITOR RESISTANT SERINE PROTEASES

The present invention relates to serine proteases of the chymotrypsin superfamily which have been modified so that they exhibit resistance to serine protease inhibitors. The invention also relates to the precursors of such compounds, their preparation, to nucleic acid coding for them and to their pharmaceutical use.

Serine proteases are endopeptidases which use serine as the nucleophile in peptide bond cleavage. There are two known superfamilies of serine proteases and these are the chymotrypsin superfamily and the Streptomyces subtilisin superfamily (Barrett, A. J., in: *Proteinase Inhibitors*, Ed. Barrett, A. J. et al., Elsevier, Amsterdam, pp 3-22 (1986) and James, M. N. G., in: *Proteolysis and Physiological Regulation*, Ed. Ribbons, D. W. et al, Academic Press, New York, pp 125-142 (1976)).

The present invention is particularly concerned with serine proteases of the chymotrypsin superfamily which includes such compounds as plasmin, tissue plasminogen activator (t-PA), urokinase-type plasminogen activator (u-PA), trypsin, chymotrypsin, granzyme, elastase, acrosin, tonin, myeloblastin, prostate-specific antigen (PSA), gamma-renin, tryptase, snake venom serine proteases, adipsin, protein C, cathepsin G, complement components C1R, C1S and C2, complement factors B, D and I, chymase, hepsin, medullasin and proteins of the blood coagulation cascade including kallikrein, thrombin, and Factors VIIa, IXa, Xa, XIa and XIIa. Members of the chymotrypsin superfamily have amino acid and structural homology of the catalytic domains, although a comparison of the sequences of the catalytic domains reveals the presence of insertions or deletions of amino acids. However, these insertions and deletions map to the surface of the folded molecule and thus do not affect the basic structure although it is likely that they contribute to the specificity of interactions of the molecule with substrates and inhibitors (Strassburger, W. et al, *FEBS Lett.*, 157, 219-223 (1983)).

Serine protease inhibitors are also well known and are divided into the following families: the bovine pancreatic trypsin inhibitor (BPTI) family, the Kazal family, the alpha-2-macroglobulin (A2M) family, the Streptomyces subtilisin inhibitor (SSI) family, the serpin family, the Kunitz family, the four-disulphide core family, the potato inhibitor family and the Bowman-Birk family.

Serine protease inhibitors inhibit their cognate serine proteases and form stable 1:1 complexes with these proteases. Structural data are available for several protease-inhibitor complexes including trypsin-BPTI, chymotrypsin-ovomucoid inhibitor and chymotrypsin-potato inhibitor (Read, R. J. et al., in: *Proteinase inhibitors*, Ed. Barrett, A. J. et al., Elsevier, Amsterdam, pp 301-336 (1986)). A structural feature which is common to all the serine protease inhibitors is a loop extending from the surface of the molecule which contains the recognition sequence for the active site of the cognate serine protease and, in fact, there is remarkable similarity in the specific interactions between different inhibitors and their cognate serine proteases, despite the diverse sequences of the inhibitors.

The serine proteases of the chymotrypsin superfamily play an important role in human and animal physiology. Some of the most important serine protease inhibitors are those which are involved in blood coagulation and fibrinolysis. In the process of blood coagulation, a cascade of enzyme activities is involved in generating a fibrin network which forms the framework of a clot or thrombus. Degra-

dation of the fibrin network (fibrinolysis) involves the protease inhibitor plasmin. Plasmin is formed in the body from its inactive precursor plasminogen by cleavage of the peptide bond between arginine 561 and valine 562 of plasminogen. This reaction is catalysed by t-PA or by u-PA.

If the balance between the clotting and fibrinolytic systems becomes locally disturbed, intravascular clots may form at inappropriate locations leading to conditions such as coronary thrombosis and myocardial infarction, deep vein thrombosis, stroke, peripheral arterial occlusion and embolism. A known way of treating such conditions is to administer to a patient a serine protease of the chymotrypsin superfamily or the precursor of such an enzyme. For example, t-PA, u-PA and plasminogen in the form of anisoylated plasminogen complexed with streptokinase are used in the treatment of myocardial infarction; plasminogen is used to supplement the natural circulatory plasminogen level to enhance thrombolytic therapy; and protein C is used as an antithrombotic agent. Serine proteases of the chymotrypsin superfamily, for example factors VIIa and IX, are administered for induction of blood clotting in disorders such as haemophilia. A major problem with the use of all of these agents in this type of therapy is their rapid neutralisation by serine protease inhibitors which reduces the efficiency of the therapy and increases the dose of agent required. It would therefore be advantageous to develop modified analogues of these endopeptidases which are resistant to inactivation by serine protease inhibitors whilst maintaining their activity. However, it is not easy to predict modifications which will result in increased resistance to inhibition without significant decrease in endopeptidase activity.

WO-A-9010649 discloses serine proteases of the chymotrypsin superfamily which have been modified and which are said to have increased resistance to serine protease inhibitors. The authors of that document have studied the known structure of the complex between trypsin and BPTI and have realised that, other than the amino acids in the major recognition site, the amino acids of trypsin that make direct contact with BPTI are located in the region between residues 37 and 41 and in the region between residues 210 to 213 of the polypeptide chain. The authors have then extrapolated from this on the basis that there is a high degree of structural homology between the catalytic domains of serine proteases and have suggested that mutation of a residue in any serine protease equivalent to the Tyr-39 residue in trypsin would lead to increased resistance of the modified analogue compared with the wild-type serine protease. They also suggest that inhibition resistant t-PA analogues can be made by mutation of an additional stretch of seven amino acids which occurs in tPA, but not in trypsin, adjacent to the predicted contact point at Arg-304 (equivalent to Tyr-39 of trypsin). However, although the catalytic domains of members of the chymotrypsin superfamily of serine proteases do, in general, have sequence and structural homology, Tyr-39 of trypsin is on a loop structure on the surface of the protein and, as is shown in FIG. 1, the equivalent regions of other serine proteases are highly variable within the superfamily. Indeed, this is acknowledged in WO-A-9010649. It is, therefore, by no means evident that the specific conformation of the loop in this region of the protein is conserved between different serine proteases, especially in cases where the number of residues in the loop differ, as is the case for trypsin and plasmin. Thus, although the residues in the region may be aligned sequentially because of the alignment of their flanking regions which do have similar sequences, it is not at all evident that their side-chains are in equivalent spatial loca-

tions and, therefore, residues which are equivalent in a sequence alignment are not necessarily able to form equivalent interactions in the folded protein. If plasmin is taken as an example, it can be seen from FIG. 1 that there are three hydrophobic residues (Phe-22, Met-24 and Phe-26) which could be involved in a similar hydrophobic interaction to that of Tyr-39 in the trypsin/BPTI complex. The numbering of the plasmin residues just mentioned is the numbering of SEQ ID No 2 which depicts the protease domain of plasmin. The residue designated 1 in SEQ ID No 2 is at position 562 of the mature protein. A study of FIG. 1 shows that any of these residues could be equivalent to Tyr-39 of trypsin which occurs at position 29 in the numbering system of FIG. 1. Clearly, therefore, the method described in WO-A-9010649 for designing a protease which is resistant to inhibition is not wholly reliable and it would be preferable to design inhibition resistant mutants in a different way.

The present inventors have realised that, because the serine protease inhibitors are structurally homologous in their active centre loop and form similar interactions with their cognate serine proteases (Read, R. J. et al., in: *Proteinase Inhibitors*, Ed. Barrett, A. J. et al., Elsevier, Amsterdam, pp 301-336 (1986)), mutations in any given serine protease which result in resistance to inhibition by a serine protease inhibitor may be applicable to mutations of spatially or sequentially equivalent residues in any other member of the chymotrypsin superfamily.

The interaction between enzyme and inhibitor responsible for inhibition of enzyme activity involves the catalytic site amino acids of the enzyme and the reactive site amino acids of the inhibitor. This principal interaction is stabilised by other interactions between the molecules. Although there is a comparatively large surface of interaction between the protease and the inhibitor, the protease/inhibitor complex is mainly stabilised by a few key interactions. These are exemplified by the interactions observed in the protease/inhibitor complex between trypsin and BPTI (Huber, R. et al., *J. Mol. Biol.* 89:73-101 (1974)), which serves as a model for the interaction between the catalytic domains of other serine proteases and their cognate inhibitors. In the trypsin/BPTI complex, the key residues of the protease, apart from those in the principal recognition site, which interact with the inhibitor are residues 37-41 and 210-213 (chymotrypsin numbering), with Tyr-39 being the most important. This interaction served as the basis for WO-A-9010649 in which the spatially equivalent residues in the t-PA/PAI-1 complex were identified, and inhibitor-resistant mutants were described.

In contrast to the disclosure WO-A-9010649, the present inventors have realised that the desired disruption of the protease/inhibitor interactions which lead to inhibitor resistance need not be caused by mutating the specific residues identified in that document or their equivalents in other serine proteases. Instead, residues in spacial, rather than sequential, proximity to these key residues, may be mutated resulting in a less stable complex between the protease and the inhibitor.

In a first aspect of the present invention, there is provided a modified endopeptidase of the chymotrypsin superfamily of serine proteases or a precursor of such an endopeptidase, which is resistant to serine protease inhibitors, characterised in that the modification comprises the mutation of one or more residues in close spacial proximity (other than sequential proximity) to a site of interaction between the protease and a cognate protease inhibitor.

In the context of this invention, the term 'precursor', when used in relation to a serine protease, refers to a protein which is cleavable by an enzyme to produce an active serine protease.

Mutations resulting in resistance to the inhibitor may induce:

- i) a conformational change in the local fold of the protease such that the resulting complex with the inhibitor is less stable than the equivalent complex between the inhibitor and the wild-type protein;
- ii) a change in the relative orientations of the protease and inhibitor on forming a complex such that the resulting complex is less stable than the equivalent complex between the inhibitor and the wild-type protein;
- iii) a change in the steric bulk of the protease in the region of the inhibitor-binding site such that the resulting complex is less stable than the equivalent complex between the inhibitor and the wild-type protein;
- iv) a change in the electrostatic potential field in the region of the inhibitor-binding site such that the resulting complex is less stable than the equivalent complex between the inhibitor and the wild-type protein; or
- v) any combination of the above.

The residues to be mutated need not be sequentially close to the key residues involved in the protease/inhibitor interaction, since the three-dimensional folding of the protease chain brings sequentially distant residues into spatial proximity. It is necessary to select the residues for mutation based on a model of either the protease used to generate the mutant, or of another member of the chymotrypsin superfamily of serine proteases. Where the three-dimensional structure of the protease to be mutated is not known, the selection of residues for mutation may be based either on a three-dimensional model of the protein to be mutated derived using homology modelling or other techniques, or on sequence alignments between the protein to be mutated and other members of the chymotrypsin superfamily of serine proteases with known three-dimensional structures. If sequence alignments are employed, it is not necessary to generate a three-dimensional structural model of the protease of interest in order to select residues for mutation to give inhibitor resistance, as spatial proximity to the key residues can be inferred from those proteins in the alignment with known three-dimensional structures. The spatial relationships between the residues to be mutated and the key residues in the protease/inhibitor interaction may be inferred by any appropriate method. Suitable methods are known to those skilled in the art.

The modified serine protease may be any serine protease of the chymotrypsin superfamily since all of these enzymes have a common mechanism of action. Examples of serine protease inhibitors which can be modified according to the present invention are as follows:

plasmin, tissue plasminogen activator (t-PA), urokinase-type plasminogen activator (u-PA), trypsin, chymotrypsin, granzyme, elastase, acrosin, tonin, myeloblastin, prostate-specific antigen (PSA), gamma-renin, tryptase, snake venom serine proteases, adipsin, protein C, cathepsin G, complement components C1R, C1S and C2, complement factors B, D and I, chymase, hepsin, medullasin and proteins of the blood coagulation cascade including kallikrein, thrombin, and Factors VIIa, IXa, Xa, XIa and XIIa.

However, modified analogues of plasmin, t-PA, u-PA, activated protein C, thrombin, factor VIIa, factor IXa, factor Xa, factor XIa and factor XIIa are particularly useful, as is a modified version of plasminogen, since all of these compounds can be used as fibrinolytic or thrombotic agents. An inhibition resistant plasmin analogue is particularly preferred.

The serine protease inhibitor to which the modified serine protease of the invention is resistant will obviously depend

on which serine protease has been modified. In the case of plasmin, the primary physiological inhibitor is $\alpha 2$ -antiplasmin which belongs to the serpin family of serine protease inhibitors. The reaction between plasmin and $\alpha 2$ -antiplasmin consists of two steps: a very fast reversible reaction between the kringle 1 lysine binding site of plasmin and the carboxy-terminal region of the inhibitor, followed by a reaction between the catalytic site of plasmin and the reactive site of the inhibitor which results in the formation of a very stable 1:1 stoichiometric enzymatically inactive complex (Holmes, W. E. et al., *J. Biol. Chem.*, 262, 1659-1664 (1987)). Therefore, when the serine protease is plasmin, it is particularly useful if the serine protease inhibitor to which the plasmin is resistant is $\alpha 2$ -antiplasmin. Plasmin is also inhibited by $\alpha 2$ -macroglobulin and $\alpha 1$ -antitrypsin and resistance to inhibition by these inhibitors is also useful.

From a three-dimensional model of the plasmin/antiplasmin complex, (described in Method 1), it has been determined that, in plasmin, the residues which are in close spatial proximity to the key residues of interaction between the protease and the inhibitor are residues 17-20, 44-54, 62, 154, 158, 198-213. The numbering used above is the numbering system of sequence ID No 2 which represents the protease domain of plasmin and begins at position 562 of the mature protein. In order to be resistant to inhibition by a serine protease inhibitor such as ($\alpha 2$ -antiplasmin, it is necessary to modify plasmin in one or more of these regions. Protease inhibition resistance can be induced in other serine proteases of the chymotrypsin superfamily by modifying equivalent regions of these proteins. FIG. 1 shows the sequences of the protease domains of a variety of proteases and, from a study of FIG. 1, it is clear where modifications should be made in order to induce resistance to protease inhibitors. In the numbering system of FIG. 1, the modification regions just mentioned occur at residues 17-22, 49-64, 72, 203, 214, and 264-281. The types of mutations which are suitable for inducing resistance to inhibition include single or multiple amino acid substitutions, additions or deletions. However, amino acid substitutions are particularly preferred.

In plasmin, examples of amino acid substitution mutations which result in a modified response to inhibition by $\alpha 2$ -antiplasmin, using the numbering system of SEQ ID No 2, are Glu-62 to Lys or Ala, Ser-17 to Leu, Arg-19 to Glu or Ala, and Glu-45 to Lys, Arg or Ala. Resistance to protease inhibition can be induced in other serine proteases by making modifications at equivalent positions. The degree of resistance to inhibition may be altered by making either single or multiple mutations in the protease, or by altering the nature of the amino acid used for substitution.

In addition to the modification of the invention, the serine protease may be modified in other ways as compared to wild-type proteins. Any modifications may be made to the protein provided that it does not lose its activity.

As an alternative to a modified serine protease, it is also possible to modify a precursor of the enzyme so that the enzyme derived from the precursor will have the desired resistance to inhibition. An example of a serine protease precursor is plasminogen which is the inactive precursor of plasmin. Conversion of plasminogen to plasmin is accomplished by cleavage of the peptide bond between arginine 561 and valine 562 of plasminogen. Under physiological conditions this cleavage is catalysed by t-PA or u-PA. Cleavage of a modified plasminogen variant of the present invention will produce a plasmin variant as described above and it is, of course, preferable that the plasminogen variant

will be cleaved to produce one of the preferred plasmin variants described above.

Again, as with serine proteases, the precursors may have other modifications. Analysis of the wild-type plasminogen molecule has revealed that it is a glycoprotein composed of a serine protease domain, five kringle domains and an N-terminal sequence of 78 amino acids which may be removed by plasmin cleavage. Cleavage by plasmin involves hydrolysis of the Arg(68)-Met(69), Lys(77)-Lys(78) or Lys(78)-Val(79) bonds to create forms of plasminogen with an N-terminal methionine, lysine or valine residue, all of which are commonly designated as lys-plasminogen. Intact plasminogen is referred to as glu-plasminogen because it has an N-terminal glutamic acid residue. Glycosylation occurs on residues Asn(289) and Thr(346) but the extent and composition are variable, leading to the presence of a number of different molecular weight forms of plasminogen in the plasma. Any of the above plasminogen variants may be modified to produce a variant according to the present invention. The protein sequencing studies of Sottrup-Jensen et al (in: *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.) 5 suppl. 3, p.95 (1978)) indicated that plasminogen was a 790 amino acid protein and that the site of cleavage was the Arg(560)-Val(561) peptide bond. A plasminogen variant which is suitable for modification according to the present invention is a 791 residue protein with an extra Ile at position 65 and encoded by cDNA isolated by Forsgren et al (*FEBS Letters*, 213, 254-260 (1987)). The serine protease domain of any of these plasminogen analogues can be recognised by its homology with serine proteases and on activation to plasmin is the catalytically active domain involved in fibrin degradation. The five kringle domains are homologous to those in other plasma proteins such as tPA and prothrombin and are involved in fibrin binding and thus localisation of plasminogen and plasmin to thrombi.

The plasminogen analogues of the present invention may also contain other modifications (as compared to wild-type glu-plasminogen) which may be one or more additions, deletions or substitutions. Examples of particularly suitable plasminogen analogues are disclosed in our copending applications WO-A-9109118 and GB 9222758.6 and comprise plasminogen analogues which are cleavable by an enzyme involved in blood clotting to produce active plasmin. These plasminogen analogues may, according to the present invention, be further modified so that, on cleavage, the plasmin which is produced is resistant to inhibition by serine protease inhibitors such as $\alpha 2$ -antiplasmin. Other plasminogen analogues which may be modified to produce the plasminogen analogues of the invention are analogues in which there has been an addition, removal, substitution or alteration of one or more kringle domains. Other suitable plasminogen analogues are Lys-plasminogen variants in which the amino terminal 68, 77 or 78 amino acids have been deleted. Such variants may have enhanced fibrin binding activity as has been observed for lys-plasminogen compared to wild-type glu-plasminogen (Bok, R. A. and Mangel, W. F., *Biochemistry*, 24, 3279-3286 (1985)). Also included within the scope of the invention are plurally-modified plasminogen analogues which include one or more modifications to prevent, reduce or alter glycosylation patterns. Such analogues may have a longer half-life, reduced plasma clearance and/or higher specific activity.

The modified serine proteases and serine protease precursors of the invention can be prepared by any suitable method and, in a second aspect of the invention, there is provided a process for the preparation of such a serine protease or serine

protease precursor, the process comprising coupling together successive amino acid residues and/or ligating oligopeptides. Although the proteins may, in principle, be synthesised wholly or partly by chemical means, it is preferred to prepare them by ribosomal translation, preferably *in vivo*, of a corresponding nucleic acid sequence. The process may further include an appropriate glycosylation step.

It is preferred to produce proteins of the invention using recombinant DNA technology. DNA encoding a naturally occurring serine protease or precursor may be obtained from a cDNA or genomic clone or may be synthesised. Amino acid substitutions, additions or deletions are preferably introduced by site-specific mutagenesis. DNA sequences encoding glu-plasminogen, lys-plasminogen, other plasminogen analogues and serine protease variants may be obtained by procedures familiar to those skilled in the art of genetic engineering.

The process for producing proteins using recombinant DNA technology will usually include the steps of inserting a suitable coding sequence into an expression vector and transfecting the vector into a suitable host cell. Therefore, in a third aspect of the invention there is provided nucleic acid coding for a modified serine protease as described above. The nucleic acid may be either DNA or RNA and may be in the form of a vector such as a plasmid, cosmid or phage. The vector may be adapted to transfect or transform prokaryotic cells, such as bacterial cells and/or eukaryotic cells, such as yeast or mammalian cells. The vector may be a cloning vector or an expression vector and comprises a cloning site and, preferably, at least one marker gene. An expression vector will additionally have a promoter operatively linked to the sequence to be inserted into the cloning site and, preferably, a sequence enabling the protein product to be secreted.

Most of the proteins of the present invention, including molecules such as tPA, can easily be obtained by inserting the coding sequence into an expression vector as described and transfecting the vector into a suitable host cell which may be a bacterium such as *E. coli*, a eukaryotic microorganism such as yeast or a higher eukaryotic cell. With molecules such as plasminogen which are unusually difficult to express, it may be necessary to use a vector of the type described in our copending application, WO-A-9109118, which comprises a first nucleic acid sequence coding for the modified serine protease, operatively linked to a second nucleic acid sequence containing a strong promoter and enhancer sequence derived from human cytomegalovirus, a third nucleic acid sequence encoding a polyadenylation sequence derived from SV40 and a fourth nucleic acid sequence coding for a selectable marker expressed from an SV40 promoter and having an additional SV40 polyadenylation signal at the 3' end of the selectable marker sequence. Such a vector may either comprise a single nucleic acid molecule or a plurality of such molecules so that, for example, the first, second and third sequences may be contained in a first nucleic acid molecule and the fourth sequence may be contained in a second nucleic acid molecule. This vector is particularly useful for the expression of plasminogen and plasminogen analogues.

For any of the proteins of the invention, the vector is preferably chosen so that the protein is expressed and secreted into the cell culture medium in a biologically active form without the need for any additional biological or chemical procedures. In the case of plasminogen, this can be achieved using the vector described above.

In a further aspect of the invention there is provided a process for the preparation of nucleic acid encoding a

modified serine protease which exhibits resistance to serine protease inhibitors, the process comprising coupling together successive nucleotides and/or ligating oligo- and/or poly-nucleotides.

In a further aspect of the invention, there is provided a cell transformed or transfected by a vector as described above. Suitable cells or cell lines include both prokaryotic and eukaryotic cells. A typical example of a eukaryotic cell is a bacterial cell such as *E. coli*. Suitable eukaryotic cells include yeast cells such as *Saccharomyces cerevisiae* or *Pichia pastoris*. Other examples of suitable eukaryotic cells are mammalian cells which grow in continuous culture and examples of such cells include Chinese hamster ovary (CHO) cells, mouse myeloma cell lines such as P3X63-Ag8.653 and NS0, COS cells, HeLa cells, 293 cells, BHK cells, melanoma cell lines such as the Bowes cell line, mouse L cells, human hepatoma cell lines such as HepG2, mouse fibroblasts and mouse NIH 3T3 cells. CHO cells are particularly suitable as hosts for the expression of plasminogen and plasminogen analogues. The transformation of the cells may be achieved by any convenient method but electroporation is a particularly suitable method.

For some molecules, such as plasminogen, there may be a low level of undesirable activation during culture. Therefore, in a further aspect of the invention, there is provided a eukaryotic host cell transfected or transformed with a first DNA sequence encoding a serpin-resistant serine protease and with an additional DNA sequence encoding the cognate inhibitor.

The modified serine proteases of the present invention have a variety of uses and, if the serine protease is a fibrinolytic or thrombolytic enzyme, it will be useful in a method for the treatment and/or prophylaxis of diseases or conditions caused by blood clotting, the method comprising administering to a patient an effective amount of the serine protease.

Therefore, in a further aspect of the invention, there is provided a modified serine protease according to the first aspect of the invention, which is a serine protease having fibrinolytic, thrombolytic, antithrombotic or prothrombotic properties, for use in medicine, particularly in the treatment of diseases mediated by blood clotting. Such conditions include myocardial and cerebral infarction, arterial and venous thrombosis, thromboembolism, post-surgical adhesions, thrombophlebitis and diabetic vasculopathies.

The invention also provides the use of a modified fibrinolytic, thrombolytic, antithrombotic or prothrombotic serine protease according to the first aspect of the invention in the preparation of an agent for the treatment and/or prophylaxis of diseases or conditions mediated by blood clotting. Examples of such conditions are mentioned above.

Furthermore, there is also provided a pharmaceutical or veterinary composition comprising one or more modified serine proteases of the first aspect of the invention together with a pharmaceutically and/or veterinarily acceptable carrier.

The composition may be adapted for administration by oral, topical or parenteral routes including intravenous or intramuscular injection or infusion. Suitable injectable compositions may comprise a preparation of the compound in isotonic physiological saline and/or buffer and may also include a local anaesthetic to alleviate the pain of the injection. Similar compositors may be used for infusions. If the compound is administered topically, it may be formulated as a cream, ointment or lotion in a suitable base.

The compounds of the invention may be supplied in unit dosage form, for example as a dry powder or water-free

concentrate in a hermetically sealed container such as an ampoule or sachet.

The quantity of material to be administered will depend on the amount of fibrinolysis or inhibition of clotting required, the required speed of action, the seriousness of the thromboembolic position and the size of the clot. The precise dose to be administered will, because of the very nature of the condition which compounds of the invention are intended to treat, be determined by the physician. As a guideline, however, a patient being treated for a mature thrombus will generally receive a daily dose of a plasminogen analogue of from 0.01 to 10 mg/kg of body weight either by injection in for example up to 5 doses or by infusion.

The invention will now be further described by way of example only with reference to the following drawings in which:

FIG. 1 shows the alignment of the catalytic domain amino acids of the chymotrypsin superfamily;

FIGS. 2a and 2b shows maps of the pGWH and pGWHgP vectors;

FIG. 3 shows the effect of $\alpha 2$ -antiplasmin on the activity of plasminogen mutant A3.

FIG. 4 shows the sequence alignment of ovalbumin and $\alpha 2$ -antiplasmin used to generate the $\alpha 2$ -antiplasmin model.

The following examples further illustrate the invention.

Examples 1 to 5 describe the expression of various plasminogen analogues from higher eukaryotic cells and example 6 describes an assay used to assess resistance to $\alpha 2$ -antiplasmin.

EXAMPLE 1

Construction and Expression of A1 and A12

The isolation of plasminogen cDNA and construction of the vectors pGWH and pGWHgP (FIG. 2) have been described in WO-A-9109118. In pGWHgP, transcription through the plasminogen cDNA can initiate at the HCMV promoter/enhancer and the selectable marker gpt is employed.

The techniques of genetic manipulation, expression and protein purification used in the manufacture of the modified plasminogen examples to follow, are well known to those skilled in the art of genetic engineering. A description of most of the techniques can be found in one of the following laboratory manuals: "Molecular Cloning" by T. Maniatis, E. F. Fritsch and J. Sambrook published by Cold Spring Harbor Laboratory, Box 100, New York, or "Basic Methods in Molecular Biology" by L. G. Davis, M. D. Digner and J. F. Battey published by Elsevier Science publishing Co Inc, New York.

Additional and modified methodologies are detailed in the methods section below.

Plasminogen analogues have been constructed which are designed to be resistant to inhibition by $\alpha 2$ -antiplasmin. A1 is a plasminogen analogue in which the amino acid Phe-587 is replaced by Asn. A12 is a plasminogen analogue in which the Arg-580 is replaced by Glu. The modification strategy in this example is essentially as described in WO-A-9109118 Example 3, with the mutagenesis reaction carried out on the 1.87 kb KpnI to HincII fragment of the thrombin activatable plasminogen analogue T19 cloned into the bacteriophage M13mp18. Single stranded template was prepared and the mutation made by oligonucleotide directed mutagenesis. For A1, a 24 base long oligonucleotide 5'GGTGCTCCA-CAATTGTGCAITTC3' (SEQ. ID. 3) was used to direct the mutagenesis and for A12 a 27 base oligonucleotide was used 5'CCAAACCTTGTTTCAAGACTGACITGC 3' (SEQ ID 7).

Plasmid DNA was introduced into CHO cells by electroporation using 800 V and 25 μ F as described in the methods section below. Selective medium (250 μ l/ml xanthine, 5 μ g/ml mycophenolic acid, 1x hypoxanthine-thymidine (HT)) was added to the cells 24 hours post transfection and the media changed every two to three days. Plates yielding gpt-resistant colonies were screened for plasminogen production using an ELISA assay. Cells producing the highest levels of antigen were re-cloned and the best producers scaled up into flasks with production being carefully monitored. Frozen stocks of all these cell lines were laid down. Producer cells were scaled up into roller bottles to provide conditioned medium from which plasminogen protein was purified using lysine SEPHAROSE 4B. (The word SEPHAROSE is a trade mark.)

EXAMPLE 2

Construction and Expression of A3 and A16

The procedure of Example 1 was generally followed except that the mutagenesis was performed on an EcoRV to HindIII fragment (0.85 kb) containing the 3' of wild type plasminogen cloned into M13. The oligonucleotide used was a 27mer 5'GTTTCGAGATTCACCTTTTGGGTGTG-CAC3' (SEQ. ID. 4) which changed Glu-623 to Lys, thus changing an acidic amino acid to a basic amino acid. The resulting mutant was cloned as an EcoRV to SphI fragment replacing the corresponding wild type sequence. The 27 base oligonucleotide 5'GTTTCGAGATTCACCTTGGGTGTG-CAC3' (SEQ ID 10) was used to change Glu-623 to Ala to produce A16.

EXAMPLE 3

Construction and Expression of A4, A14 and A15

Mutant A4 is designed to disrupt ionic interactions on the surface of plasminogen preventing binding to antiplasmin. The mutagenesis and sub-cloning strategy was as described in Example 1 using a 24 base oligonucleotide 5'CTTGGG-GACTTCTTCAAGCAGTGG3' (SEQ. ID. 5) designed to convert Glu-606 to Lys. The 24 base oligonucleotide 5'CTTGGGGACTTGGCTAGACAGTGG 3' (SEQ ID 8) was used to change Glu-606 to Ala to produce A14 and the 25 base oligonucleotide 5'CTTGGGGACTTCCTTAGA-CAGTGGG 3' (SEQ ID 9) was used to change Glu-606 to Arg to produce A15.

EXAMPLE 4

Construction and Expression of A5

Plasminogen analogue A5 was designed to alter the positioning of the Tyr 39 containing structural loop and was made generally as described in the procedure of Example 1. In A5, Ser-578 has been replaced by Leu using the 24mer 5'CTCGTACGAAGCAGGACTTGCCAG3' (SEQ. ID. 6) on the KpnI to EcoRV fragment of plasminogen in M13 as the template. The mutation was cloned directly into pGW1Hg.plasminogen using the restriction enzymes HindIII and SphI. These sites had previously been introduced at the extreme 5' end of plasminogen and at 1850 respectively via mutagenesis; the plasminogen coding sequence was not affected by this procedure.

EXAMPLE 5

Construction and Expression of double mutant A3A4

Plasminogen mutant A3A4 combines the two mutations A3 and A4 as described in Examples 2 and 3 respectively.

Mutagenesis was performed on the EcoRV to SphI fragment of A4 cloned into M13 using the A3 mutagenesis oligonucleotide (SEQ ID4).

EXAMPLE 6

Plasmin-Antiplasmin Interaction Assays

A chromogenic assay was used to assess the resistance of the plasmin(ogen) mutants to inhibition by $\alpha 2$ -antiplasmin. Inhibition of plasmin activity was determined by the change in the rate of cleavage of the plasmin chromogenic substrate S2251 (Quadrach, P.O. Box 167, Epsom, Surrey, KT17 2SB).

Prior to assay, the plasminogens were activated to plasmin using either urokinase for mutants in wild type plasminogen, or thrombin for thrombin activatable plasminogen mutants (WO-A-9109118). Activation of wild-type plasminogen to plasmin was achieved by incubation of the plasminogen (ca. 14 μ g) with urokinase (16.8×10^{-3} U) in 1750 μ l of assay buffer (50 mM Tris, 0.1 mM EDTA, 0.00005% Triton X100, 0.1% (w/v) human serum albumin, pH 8.0) at 37° C. for 5 mins. Activation of thrombin activatable plasminogen mutants to plasmin was achieved by incubation of the plasminogen (ca. 14 μ g) with thrombin in 1750 μ l of assay buffer at 37° C. Hirudin was added to inhibit the thrombin activity as thrombin cleaves the chromogenic substrate.

Plasmin (125 μ l) was mixed with 250 μ l S2251 (2 mg/ml in assay buffer) and 125 μ l antiplasmin (1.25 μ g in assay buffer, #4032 American Diagnostica Inc., 222 Railroad Avenue, P.O. Box 1165, Greenwich, Conn. 06836-1165) or 125 μ l assay buffer in a cuvette and the absorbance at 405 nM measured over time.

A Beckman DU64 spectrophotometer and Beckman "Data Leader" data capture software were used to record absorbance at 405 nM at 1 sec intervals for 8 minutes. The Data Leader software package was used to calculate the first derivative of the data to provide the rate of change of absorbance at 405nm against time, an estimate of active plasmin concentration against time. Wild type plasmin was rapidly inactivated by $\alpha 2$ -antiplasmin; after only 15 seconds the plasmin was essentially inactivated. In contrast, plasminogen mutant A3 has an antiplasmin resistant phenotype and is only slowly inactivated by antiplasmin with a $t_{1/2}$ (half the rate of OD change at $t=15$ sec) of approximately 75 seconds (FIG. 3).

METHODS

1. Model structures were built by homology based on the x-ray structures of trypsin/BPTI. A refined plasminogen structure was modelled by homology to thrombin using the PPACK/thrombin x-ray structure from Bode et al. (Bode, W. et al., *EMBO J.* 8:3467-3475 (1989)). A refined $\alpha 2$ -antiplasmin [A2AP] structure was modelled by homology to ovalbumin using atomic co-ordinates from the Brookhaven Protein Data Bank entry 1OVA, except for the loop containing the reactive bond, which was modelled using the co-ordinates for residues 13 to 19 of BPTI from the PDB entry 2PTC. The alignment used to generate the A2AP model is shown in FIG. 4. The A2AP model described here does not include co-ordinates for the 79 N-terminal residues and 55 C-terminal residues.

Most serine-protease-directed inhibitors react with cognate enzymes according to a common, substrate-like standard mechanism (Bode, W. and Huber, R., *Eur. J. Biochem.* 204:433-451 (1992)). In particular, they all possess an exposed active site-binding loop with a characteristic

canonical conformation. The binding loop on the A2AP model was therefore modelled on the equivalent loop of BPTI (residues 13 to 19), using atomic co-ordinates from the PDB entry 2PTC (in which BPTI is complexed with trypsin).

The complex of A2AP and the plasmin serine protease domain was modelled using the trypsin/BPTI complex structure from PDB entry 2PTC. The A2AP model was fitted to the BPTI structure by optimising the RMS difference between the co-ordinates of the backbone atoms in the active site-binding loops of the two inhibitors. The plasmin serine protease domain model was fitted to the trypsin structure by optimising the RMS difference between the co-ordinates of the C-alpha atoms of the conserved residues in an optimal sequence alignment of the two proteins. The A2AP/plasmin complex model was then refined by energy-minimisation.

The homology modelling was performed on a Silicon Graphics Indigo workstation using the Quanta molecular modelling program from Molecular Simulations Incorporated. Sequence alignments were produced using Quanta, the GCG sequence analysis software from the University of Wisconsin (Devereux, Haeblerli and Smithies, *Nucleic Acids Research* 12(1):387-395 (1984), and proprietary sequence alignment software. However, the actual method by which the homology models were built is not critical to this invention.

The trypsin and BPTI sequences used in the homology modelling were obtained from the Brookhaven Protein Data Bank atomic co-ordinate entry 2PTC, the thrombin sequence was obtained from the PPACK/thrombin co-ordinate file, the plasminogen sequence from the SWISSPROT database entry PLMN_HUMAN, and the A2AP sequence from the SWISSPROT entry A2AP_HUMAN.

2. Mung Bean Nuclease Digestion

10 units of mung bean nuclease was added to approximately 1 μ g DNA which had been digested with a restriction enzyme in a buffer containing 30 mM NaOAc pH5.0, 100 mM NaCl, 2 mM ZnCl₂, 10% glycerol. The mung bean nuclease was incubated at 37° for 30 minutes, inactivated for 15 minutes at 67° before being phenol extracted and ethanol precipitated.

3. Oligonucleotide synthesis

The oligonucleotides were synthesised by automated phosphoramidite chemistry using cyanoethyl phosphoramidites. The methodology is now widely used and has been described (Beaucage, S. L. and Caruthers, M. H. *Tetrahedron Letters* 24, 245 (1981) and Caruthers, M. H. *Science* 230, 281-285 (1985)).

4. Purification of Oligonucleotides

The oligonucleotides were de-protected and removed from the CPG support by incubation in concentrated NH₃. Typically, 50 mg of CPG carrying 1 micromole of oligonucleotide was de-protected by incubation for 5 hours at 70° in 600 μ l of concentrated NH₃. The supernatant was transferred to a fresh tube and the oligomer precipitated with 3 volumes of ethanol. Following centrifugation the pellet was dried and resuspended in 1 ml of water. The concentration of crude oligomer was then determined by measuring the absorbance at 260 nm. For gel purification 10 absorbance units of the crude oligonucleotide was dried down and resuspended in 15 μ l of marker dye (90% de-ionised formamide, 10 mM tris, 10 mM borate, 1 mM EDTA, 0.1% bromophenol blue). The samples were heated at 90° for 1 minute and then loaded onto a 1.2 mm thick denaturing polyacrylamide gel with 1.6 mm wide slots. The gel was prepared from a stock of 15% acrylamide, 0.6% bisacrylamide and 7M urea in 1X TBE and was polymerised with

0.1% ammonium persulphate and 0.025% TEMED. The gel was pre-run for 1 hr. The samples were run at 1500 V for 4–5 hours. The bands were visualised by UV shadowing and those corresponding to the full length product cut out and transferred to micro-testubes. The oligomers were eluted from the gel slice by soaking in AGEB (0.5M ammonium acetate, 0.01M magnesium acetate and 0.1% SDS) overnight. The AGEB buffer was then transferred to fresh tubes and the oligomer precipitated with three volumes of ethanol at 70° for 15 mins. The precipitate was collected by centrifugation in an Eppendorf microfuge for 10 mins, the pellet washed in 80% ethanol, the purified oligomer dried, redissolved in 1 ml of water and finally filtered through a 0.45 micron micro-filter. (The word EPPENDORF is a trade mark.) The concentration of purified product was measured by determining its absorbance at 260 nm.

5. Kinasing of Oligomers

100 pmole of oligomer was dried down and resuspended in 20 µl kinase buffer (70 mM Tris pH 7.6, 10 mM MgCl₂, 1 mM ATP, 0.2 mM spermidine, 0.5 mM dithiothreitol). 10 u of T4 polynucleotide kinase was added and the mixture incubated at 37° for 30 mins. The kinase was then inactivated by heating at 70° for 10 mins.

6. Dideoxy Sequencing

The protocol used was essentially as has been described (Biggin, M. D., Gibson, T. J., Hong, G. F. P.N.A.S. 80 3963–3965 (1983). Where appropriate the method was modified to allow sequencing on plasmid DNA as has been described (Guo, L-H., Wu R Nucleic Acids Research 11 5521–5540 (1983).

7. Transformation

Transformation was accomplished using standard procedures. The strain used as a recipient in the cloning using plasmid vectors was HW87 or DH5 which has the following genotype:

araD139(ara-leu)del7697 (lacI^{POZY})del74 galU galK hsdR rpsL sr1 recA56

RZ1032 is a derivative of *E. coli* that lacks two enzymes of DNA metabolism: (a) dUTPase (dut) which results in a high concentration of intracellular dUTP, and (b) uracil N-glycosylase (ung) which is responsible for removing misincorporated uracils from DNA (Kunkel et al, Methods in Enzymol., 154, 367–382 (1987)). its principal benefit is that these mutations lead to a higher frequency of mutants in site directed mutagenesis. RZ1032 has the following genotype:

HfrKL16PO/45[lysA961-62], dut1, ung1, thi1, re[A], Zhd-279::Tn10, supE44

JM103 is a standard recipient strain for manipulations involving M13 based vectors.

8. Site Directed Mutagenesis

Kinased mutagenesis primer (2.5 pmole) was annealed to the single stranded template DNA, which was prepared using RZ1032 as host, (1 µg) in a final reaction mix of 10 µl containing 70 mM Tris, 10 mM MgCl₂. The reaction mixture in a polypropylene micro-testtube (EPPENDORF) was placed in a beaker containing 250 ml of water at 70° C. for 3 minutes followed by 37° C. for 30 minutes. The annealed mixture was then placed on ice and the following reagents added: 1 µl of 10 X TM (700 mM Tris, 100 mM MgCl₂ pH7.6), 1 µl of a mixture of all 4 deoxyribonucleotide triphosphates each at 5 mM, 2 µl of T4 DNA ligase (100u), 0.5 µl Klenow fragment of DNA polymerase and 4.5 µl of water. The polymerase reaction mixture was then incubated

at 15° for 4–16 hrs. After the reaction was complete, 180 µl of TE (10 mM Tris, 1 mM EDTA pH8.0) was added and the mutagenesis mixture stored at –20° C. For the isolation of mutant clones the mixture was then transformed into the recipient JM103 as follows. A 5 ml overnight culture of JM103 in 2 X YT (1.6% Bactotryptone, 1% Yeast Extract, 1% NaCl) was diluted 1 in a 100 into 50 ml of pre-warmed 2 X YT. The culture was grown at 37° with aeration until the A600 reached 0.4. The cells were pelleted and resuspended in 0.5 vol of 50 mM CaCl₂ and kept on ice for 15 mins. The cells were then re-pelleted at 4° and resuspended in 2.5 ml cold 50 mM CaCl₂. For the transfection, 0.25, 1, 2, 5, 20 and 50 µl aliquots of the mutagenesis mixture were added to 200 µl of competent cells which were kept on ice for 30 mins. The cells were then heated shocked at 42° for 2 mins. To each tube was then added 3.5 ml of YT soft agar containing 0.2 ml of a late exponential culture of JM103, the contents were mixed briefly and then poured onto the surface of a pre-warmed plate containing 2 X YT solidified with 1.5% agar. The soft agar layer was allowed to set and the plates then incubated at 37° overnight.

Single stranded DNA was then prepared from isolated clone as follows: Single plaques were picked into 4 ml of 2 X YT that had been seeded with 10 µl of a fresh overnight culture of JM103 in 2 X YT. The culture was shaken vigorously for 6 hrs. 0.5 ml of the culture was then removed and added to 0.5 ml of 50% glycerol to give a reference stock that was stored at –20°. The remaining culture was centrifuged to remove the cells and 1 ml of supernatant carrying the phage particles was transferred to a fresh EPPENDORF tube. 250 µl of 20% PEG6000, 250 mM NaCl was then added, mixed and the tubes incubated on ice for 15 mins. The phage were then pelleted at 10,000 rpm for 10 mins, the supernatant discarded and the tubes re-centrifuged to collect the final traces of PEG solution which could then be removed and discarded. The phage pellet was thoroughly resuspended in 200 µl of TEN (10 mM Tris, 1 mM EDTA, 0.3M NaOAc). The DNA was isolated by extraction with an equal volume of Tris saturated phenol. The phases were separated by a brief centrifugation and the aqueous phase transferred to a clean tube. The DNA was re-extracted with a mixture of 100 µl of phenol, 100 µl chloroform and the phases again separated by centrifugation. Traces of phenol were removed by three subsequent extractions with chloroform and the DNA finally isolated by precipitation with 2.5 volumes of ethanol at –20° overnight. The DNA was pelleted at 10,000 rpm for 10 min, washed in 70% ethanol, dried and finally resuspended in 50 µl of TE.

9. Electroporation

Chinese hamster ovary cells (CHO) or the mouse myeloma cell line p3×63-Ag8.653 were grown and harvested in mid log growth phase. The cells were washed and resuspended in PBS and a viable cell count was made. The cells were then pelleted and resuspended at 1×10⁷ cells/ml. 40 µg of linearised DNA was added to 1 ml of cells and allowed to stand on ice for 15 mins. One pulse of 800 V/ 25 µF was administered to the cells using a commercially available electroporation apparatus (BIORAD GENE PULSER—trade mark). The cells were incubated on ice for a further 15 mins and then plated into 5 ×96 well plates with 200 µl of medium per well (DMEM, 5% FCS, Pen/Strep, glutamine) or 3×9 cm dishes with 10 mls medium in each dish and incubated overnight. After 24 hrs the medium was removed and replaced with selective media containing xanthine (250 µg/ml), mycophenolic acid (5 µg/ml) and 1×hypoxanthine-thymidine (HT). The cells were fed every third day. After about 14 days gpt resistant colonies are

15

evident in some of the wells and on the plates. The plates were screened for plasminogen by removing an aliquot of medium from each well or plate and assayed using an ELISA assay. Clones producing plasminogen were scaled up and the expression level monitored to allow the selection of the best producer.

10. ELISA for Human Plasminogen

ELISA plates (Pro-Bind, Falcon) are coated with 50 µl/well of goat anti-human plasminogen serum (Sigma) diluted 1:1000 in coating buffer (4.0 g Na₂CO₃(10.H₂O), 2.93 g NaHCO₃ per liter H₂O, pH 9.6) and incubated overnight at 4° C. Coating solution is then removed and plates are blocked by incubating with 50 µl/well of PBS/0.1% casein at room temperature for 15 minutes. Plates are then washed 3 times with PBS/0.05% Tween 20. Samples of plasminogen or standards diluted in PBS/Tween are added to the plate and incubated at room temperature for 2 hours. The plates are then washed 3 times with PBS/Tween and then 50 µl/well of a 1:1000 dilution in PBS/Tween of a monoclonal antihuman plasminogen antibody (eg #3641 and #3642 from American Diagnostica, New York, U.S.A.) is added and incubated at room temperature for 1 hour. The plates are again washed 3 times with PBS/Tween and then 50 µl/well of horse radish peroxidase conjugated goat anti-mouse IgG (Sigma) is added and incubated at room temperature for 1 hour. Alternatively, the bound plasminogen is revealed by incubation with 50 µl/well of horse radish peroxidase conjugated sheep anti-human plasminogen (The Binding Site). The plates are washed 5 times with PBS/Tween and then incubated with 100 µl/well of peroxidase substrate (0.1M sodium acetate/citric acid buffer pH 6.0 containing 100

16

mg/liter 3,3',5,5'-tetramethyl benzidine and 13 mM H₂O₂. The reaction is stopped after approximately 5 minutes by the addition of 25 µl/well of 2.5M sulphuric acid and the absorbance at 450 nm read on a platereader.

11. Purification of Plasminogen Variants

Plasminogen variants are purified in a single step by chromatography on lysine SEPHAROSE 4B (Pharmacia). A column is equilibrated with at least 10 column volumes of 0.05M sodium phosphate buffer pH 7.5. The column is loaded with conditioned medium at a ratio of 1 ml resin per 0.6 mg of plasminogen variant as determined by ELISA using human glu-plasminogen as standard. Typically 400 ml of conditioned medium containing plasminogen are applied to a 10 ml column (H:D=4) at a linear flow rate of 56 ml/cm/h at 4° C. After loading is complete, the column is washed with a minimum of 5 column volumes of 0.05M phosphate buffer pH 7.5 containing 0.5M NaCl until non-specifically bound protein ceases to be eluted. Desorption of bound plasminogen is achieved by the application of 0.2M epsilon-amino-caproic acid in de-ionised water pH 7.0. Elution requires 2 column volumes and is carried out at a linear flow rate of 17 ml/cm/h. Following analysis by SDS PAGE to check 10 purity, epsilon-amino-caproic acid is subsequently removed and replaced with a suitable buffer, eg Tris, PBS, HEPES or acetate, by chromatography on pre-packed, disposable, PD10 columns containing SEPHADEX G-25M (Pharmacia (The word SEPHADEX is a trade mark.) Typically, 2.5 ml of each plasminogen mutant at a concentration of 0.3 mg/ml are processed in accordance with the manufacturers' instructions. Fractions containing plasminogen, as determined by A280 are then pooled.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 10

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 690 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i i i) HYPOTHETICAL: NO

(i v) ANTI-SENSE: NO

(v i) ORIGINAL SOURCE:

(A) ORGANISM: Homo sapiens

(i x) FEATURE:

- (A) NAME/KEY: CDS
- (B) LOCATION: 1..690
- (D) OTHER INFORMATION: /partial
- / codon_start=1
- / function="encodes plasmin protease domain"
- / product="nucleotide with corresponding protein"
- / number=1

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
GTT GTA GGG GGG TGT GTG GCC CAC CCA CAT TCC TGG CCC TGG CAA GTC
Val Val Gly Gly Cys Val Ala His Pro His Ser Trp Pro Trp Glu Val
1          5          10          15
```

5,645,833

17

18

-continued

AGT CTT AGA ACA AGG TTT GGA ATG CAC TTC TGT GOA GGC ACC TTG ATA	96
Ser Leu Arg Thr Arg Phe Gly Met His Phe Cys Gly Gly Thr Leu Ile	
20 25 30	
TCC CCA GAG TGG GTG TTG ACT GCT GCC CAC TGC TTG GAG AAG TCC CCA	144
Ser Pro Glu Trp Val Leu Thr Ala Ala His Cys Leu Glu Lys Ser Pro	
35 40 45	
AGG CCT TCA TCC TAC AAG GTC ATC CTG GGT GCA CAC CAA GAA GTG AAT	192
Arg Pro Ser Ser Tyr Lys Val Ile Leu Gly Ala His Gln Glu Val Asn	
50 55 60	
CTC GAA CCG CAT GGT CAG GAA ATA GAA GTG TCT AGG CTG TTC TTG GAG	240
Leu Glu Pro His Gly Gln Glu Ile Glu Val Ser Arg Leu Phe Leu Glu	
65 70 75 80	
CCC ACA CGA AAA GAT ATT GCC TTG CTA AAG CTA AGC AGT CCT GCC GTC	288
Pro Thr Arg Lys Asp Ile Ala Leu Leu Lys Leu Ser Ser Pro Ala Val	
85 90 95	
ATC ACT GAC AAA GTA ATC CCA GCT TGT CTG CCA TCC CCA AAT TAT GTG	336
Ile Thr Asp Lys Val Ile Pro Ala Cys Leu Pro Ser Pro Asn Tyr Val	
100 105 110	
GTC GCT GAC CGG ACC GAA TGT TTC ATC ACT GGC TGG GGA GAA ACC CAA	384
Val Ala Asp Arg Thr Glu Cys Phe Ile Thr Gly Trp Gly Glu Thr Gln	
115 120 125	
GGT ACT TTT GGA GCT GGC CTT CTC AAG GAA GCC CAG CTC CCT GTG ATT	432
Gly Thr Phe Gly Ala Gly Leu Leu Lys Glu Ala Gln Leu Pro Val Ile	
130 135 140	
GAG AAT AAA GTG TGC AAT CGC TAT GAG TTT CTG AAT GGA AGA GTC CAA	480
Glu Asn Lys Val Cys Asn Arg Tyr Glu Phe Leu Asn Gly Arg Val Gln	
145 150 155 160	
TCC ACC GAA CTC TGT GCT GGG CAT TTG GCC GGA GGC ACT GAC AGT TGC	528
Ser Thr Glu Leu Cys Ala Gly His Leu Ala Gly Gly Thr Asp Ser Cys	
165 170 175	
CAG GGT GAC AGT GGA GGT CCT CTG GTT TGC TTC GAG AAG GAC AAA TAC	576
Gln Gly Asp Ser Gly Gly Pro Leu Val Cys Phe Glu Lys Asp Lys Tyr	
180 185 190	
ATT TTA CAA GGA GTC ACT TCT TGG GGT CTT GGC TGT GCA CGC CCC AAT	624
Ile Leu Gln Gly Val Thr Ser Trp Gly Leu Gly Cys Ala Arg Pro Asn	
195 200 205	
AAG CCT GGT GTC TAT GTT CGT GTT TCA AGG TTT GTT ACT TGG ATT GAG	672
Lys Pro Gly Val Tyr Val Arg Val Ser Arg Phe Val Thr Trp Ile Glu	
210 215 220	
GGA GTG ATG AGA AAT AAT	690
Gly Val Met Arg Asn Asn	
225 230	

(2) INFORMATION FOR SEQ ID NO2:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 230 amino acids

(B) TYPE: amino acid

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO2:

Val Val Gly Gly Cys Val Ala His Pro His Ser Trp Pro Trp Gln Val	
1 5 10 15	
Ser Leu Arg Thr Arg Phe Gly Met His Phe Cys Gly Gly Thr Leu Ile	
20 25 30	
Ser Pro Glu Trp Val Leu Thr Ala Ala His Cys Leu Glu Lys Ser Pro	
35 40 45	
Arg Pro Ser Ser Tyr Lys Val Ile Leu Gly Ala His Gln Glu Val Asn	
50 55 60	

-continued

```

Leu Glu Pro His Gly Gln Glu Ile Glu Val Ser Arg Leu Phe Leu Glu
 65          70          75          80
Pro Thr Arg Lys Asp Ile Ala Leu Leu Lys Leu Ser Ser Pro Ala Val
          85          90          95
Ile Thr Asp Lys Val Ile Pro Ala Cys Leu Pro Ser Pro Asn Tyr Val
          100          105          110
Val Ala Asp Arg Thr Glu Cys Phe Ile Thr Gly Trp Gly Glu Thr Gln
          115          120          125
Gly Thr Phe Gly Ala Gly Leu Leu Lys Glu Ala Gln Leu Pro Val Ile
          130          135          140
Glu Asn Lys Val Cys Asn Arg Tyr Glu Phe Leu Asn Gly Arg Val Gln
          145          150          155
Ser Thr Glu Leu Cys Ala Gly His Leu Ala Gly Gly Thr Asp Ser Cys
          165          170          175
Gln Gly Asp Ser Gly Gly Pro Leu Val Cys Phe Glu Lys Asp Lys Tyr
          180          185          190
Ile Leu Gln Gly Val Thr Ser Trp Gly Leu Gly Cys Ala Arg Pro Asn
          195          200          205
Lys Pro Gly Val Tyr Val Arg Val Ser Arg Phe Val Thr Trp Ile Glu
          210          215          220
Gly Val Met Arg Asn Asn
          225          230

```

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i i i) HYPOTHETICAL: NO

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..24
- (D) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
FOR A1"
/ product="SYNTHETIC DNA"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GGTGCCTCCA CAATTGTGCA TTCC

24

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 27 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..27
- (D) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
FOR A3"
/ product="SYNTHETIC DNA"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

GTTCGAAGATT CACTTTTTTGG TGTGCAC

27

-continued

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..24
- (D) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
FOR A4"
/ product="SYNTHETIC DNA"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

CTTGGGGACT TCTTCAAGCA GTGG

2 4

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..24
- (D) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
USED FOR A5"
/ product="SYNTHETIC DNA"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CTCGTACGAA GCAGGACTTG CCA G

2 4

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 27 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..27
- (D) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
FOR A12"
/ product="SYNTHETIC DNA"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

CCAAACCTTG TTTCAAGACT GACTTGC

2 7

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 24 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..24

-continued

```

(D ) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
    FOR A14"
    / product="SYNTHETIC DNA"

(x i ) SEQUENCE DESCRIPTION: SEQ ID NO:8:

CTTGGGGACT TGGCTAGACA GTGG                                     2 4

( 2 ) INFORMATION FOR SEQ ID NO:9:

    ( i ) SEQUENCE CHARACTERISTICS:
        ( A ) LENGTH: 25 base pairs
        ( B ) TYPE: nucleic acid
        ( C ) STRANDEDNESS: single
        ( D ) TOPOLOGY: linear

    ( i i ) MOLECULE TYPE: cDNA

    ( i x ) FEATURE:
        ( A ) NAME/KEY: misc_feature
        ( B ) LOCATION: 1..25
        ( D ) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
            FOR A15"
            / product="SYNTHETIC DNA"

    ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:9:

CTTGGGGACT TCCTTAGACA GTGGG                                     2 5

( 2 ) INFORMATION FOR SEQ ID NO:10:

    ( i ) SEQUENCE CHARACTERISTICS:
        ( A ) LENGTH: 27 base pairs
        ( B ) TYPE: nucleic acid
        ( C ) STRANDEDNESS: single
        ( D ) TOPOLOGY: linear

    ( i i ) MOLECULE TYPE: synthetic DNA

    ( i x ) FEATURE:
        ( A ) NAME/KEY: misc_feature
        ( B ) LOCATION: 1..27
        ( C ) OTHER INFORMATION: /function="MUTAGENESIS PRIMER
            FOR A16"
            / product="SYNTHETIC DNA"

    ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GTTGAGATT CACTGCTTGG TGTGCAC                                     2 7

```

We claim:

1. A plasmin modified so as to exhibit resistance to inhibitors of plasmin, characterized in that the modification comprises the mutation of the residue in a region corresponding to residue 17 according to the numbering of SEQ ID NO 2.

2. A plasmin modified so as to exhibit resistance to inhibitors of plasmin, characterized in that the modification comprises the mutation of one or more residues in a region corresponding to residues 44 to 54 according to the numbering of SEQ ID NO 2.

3. A plasmin modified so as to exhibit resistance to inhibitors of plasmin, characterized in that the modification comprises the mutation of the residue in a region corresponding to residue 45 according to the numbering of SEQ ID NO 2.

4. A plasmin modified so as to exhibit resistance to inhibitors of plasmin, characterized in that the modification comprises the mutation of the residue in a region corresponding to residue 62 according to the numbering of SEQ ID NO 2.

5. A plasmin modified so as to exhibit resistance to inhibitors of plasmin, characterized in that the modification

comprises the mutation of one or more residues in a region corresponding to residues 202 or 203 according to the numbering of SEQ ID NO 2.

6. A plasmin modified so as to exhibit resistance to inhibitors of plasmin, characterized in that the modification comprises the mutation of one or more residues in a region or regions corresponding to residues 17, 44 to 54, 62, 202 and 203, according to the numbering of SEQ ID NO 2.

7. A plasmin as claimed in claim 6, which has one or more of the following mutations: Ser-17 to Leu, Glu-45 to Lys or Arg, or Glu-62 to Lys or Ala, according to the numbering of SEQ ID NO 2.

8. A plasmin as claimed in claim 7, which has the following mutations: Glu-62 to Lys and Glu-45 to Lys, according to the number of SEQ ID NO 2.

9. A plasmin precursor, which, when cleaved, forms a plasmin modified so as to exhibit resistance to inhibitors of plasmin, characterized in that the modification comprises the mutation of one or more residues in a region or regions corresponding to residues 17, 44 to 54, 62, 202, and 203, according to the numbering of SEQ ID NO 2.

10. A plasmin precursor, which, when cleaved, forms a modified plasmin as claimed in claim 7.

25

11. A plasmin precursor, which, when cleaved, forms said modified plasmin of claim 8.

12. An isolated nucleotide sequence coding for said plasmin precursor of claim 9.

13. The isolated nucleotide sequence of claim 12, further comprising a first nucleic acid sequence coding for said modified plasmin, operatively linked to a second nucleic acid sequence containing a strong promoter and enhancer sequence derived from human cytomegalovirus, a third nucleic acid sequence encoding a polyadenylation sequence 5
10 derived from SV40 and a fourth nucleic acid sequence coding for a selectable marker expressed from an SV40 promoter and having an additional SV40 polyadenylation signal at the 3' end of the selectable marker sequence.

14. An expression vector comprising the nucleic acid 15 sequence as in claims 12 or 13.

15. The vector of claim 14, wherein said vector is selected from the group consisting of a plasmid, a cosmid, and a phage.

16. A cell transformed or transfected with the expression 20 vector of claim 14.

26

17. The cell of claim 16, wherein said cell is additionally transfected or transformed by an expression vector comprising a nucleic acid sequence coding for a plasmin inhibitor.

18. The cell of claim 17, wherein said plasmin inhibitor is selected from the group consisting of alpha2-antiplasmin, alpha2-macroglobulin and alpha1-antitrypsin.

19. A pharmaceutical composition comprising a modified plasmin as claimed in any one of claims 1 to 8, together with a pharmaceutically acceptable carrier.

20. A pharmaceutical composition comprising a modified plasmin precursor as claimed in any one of claims 9 to 11, together with a pharmaceutically acceptable carrier.

21. A veterinary composition for use in mammals, comprising a modified plasmin as claimed in any one of claims 1 to 8, together with a carrier acceptable for veterinary use.

22. A veterinary composition for use in mammals, comprising a modified plasmin precursor as claimed in any one of claims 9 to 11, together with a carrier acceptable for veterinary use.

* * * * *

UNITED STATES PATENT AND TRADEMARK OFFICE
CERTIFICATE OF CORRECTION

PATENT NO. : 5,645,833

DATED : July 8, 1997

Page 1 of 2

INVENTOR(S) : Keith Martyn Dawson and Richard James Gilbert

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

In claim 9, at column 24, line 64, after "44 to 54," insert — and — .

In claim 13, at column 25, line 7, after "modified plasmin" insert — precursor —.

In claim 19, at column 26, line 8, after "claims 1" insert -- to 4, and 6 --.

In claim 21, at column 21, line 15, after "claims 1" insert -- to 4, and 6 --.

UNITED STATES PATENT AND TRADEMARK OFFICE
CERTIFICATE OF CORRECTION

PATENT NO. : 5,645,833

Page 2 of 2

DATED : July 8, 1997

INVENTOR(S) : Keith Martyn Dawson and Richard James Gilbert

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

At column 23, lines 66 to 67 and column 24, lines 45 to 47, cancel claim 5.

In claim 6, at column 24, lines 51 to 52, cancel "202 and 203".

In claim 6, at column 24, line 51, after "44 to 54," insert — and — .

In claim 9, at column 24, line 64, cancel "202, and 203".

Signed and Sealed this

Third Day of February, 1998

Attest:



BRUCE LEHMAN

Attesting Officer

Commissioner of Patents and Trademarks

UNITED STATES PATENT AND TRADEMARK OFFICE
CERTIFICATE OF CORRECTION

PATENT NO. : 5,645,833

DATED : July 8, 1997

Page 1 of 2

INVENTOR(S) : Keith Martyn Dawson and Richard James Gilbert

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

In claim 9, at column 24, line 64, after "44 to 54," insert — and — .

In claim 13, at column 25, line 7, after "modified plasmin" insert — precursor —.

In claim 19, at column 26, line 8, after "claims 1" insert -- to 4, and 6 --.

In claim 21, at column 21, line 15, after "claims 1" insert -- to 4, and 6 --.

UNITED STATES PATENT AND TRADEMARK OFFICE
CERTIFICATE OF CORRECTION

PATENT NO. : 5,645,833

Page 2 of 2

DATED : July 8, 1997

INVENTOR(S) : Keith Martyn Dawson and Richard James Gilbert

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

At column 23, lines 66 to 67 and column 24, lines 45 to 47, cancel claim 5.

In claim 6, at column 24, lines 51 to 52, cancel "202 and 203".

In claim 6, at column 24, line 51, after "44 to 54," insert — and — .

In claim 9, at column 24, line 64, cancel "202, and 203".

Signed and Sealed this

Third Day of February, 1998

Attest:



BRUCE LEHMAN

Attesting Officer

Commissioner of Patents and Trademarks



Exhibit 12

A comprehensive set of sequence analysis programs for the VAX

John Devereux, Paul Haeblerli* and Oliver Smithies

Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

Received 18 August 1983

ABSTRACT

The University of Wisconsin Genetics Computer Group (UWGCG) has been organized to develop computational tools for the analysis and publication of biological sequence data. A group of programs that will interact with each other has been developed for the Digital Equipment Corporation VAX computer using the VMS operating system. The programs available and the conditions for transfer are described.

INTRODUCTION

The rapid advances in the field of molecular genetics and DNA sequencing have made it imperative for many laboratories to use computers to analyze and manage sequence data. UWGCG was founded when it became clear to several faculty members at the University of Wisconsin that there was no set of sequence analysis programs that could be used together as a coherent system and be modified easily in response to new ideas.

With intramural support a computer group was organized to build a strong foundation of software upon which future programs in molecular genetics could be based. This initial project has been completed and the resulting programs, written in Fortran 77, are available for VAX computers using the VMS operating system. Most of the programs can be used with only a terminal, although several require a Hewlett Packard plotter.

UWGCG software has been installed for testing at eight different institutions. A simple method has been developed for transferring and maintaining this system on other VAX computers.

DESIGN PRINCIPLES

UWGCG program design is based on the "software tools" approach of Kernighan and Plauger(1). Each program performs a simple function and is easy to use. The programs can be used independently in different combinations so

Nucleic Acids Research

that complex problems are solved by the use of several programs in succession. New programming is simplified since less effort is required to bridge a gap between existing programs.

UWGCG software is designed to be maintained and modified at sites other than the University of Wisconsin. The program manual is extensive and the source codes are organized to make modification convenient. Scientists using UWGCG software are encouraged to use existing programs as a framework for developing new ones. Our copyright can be removed from any program modified by more than 25% of our original effort.

PROGRAMS AVAILABLE FROM UWGCG

The programs described below are named and defined individually in Table 1. Program names in the text are underlined.

Comparisons

Comparisons may be done with "dot plots" using the method of Maizel and Lenk(2). Optimal alignments can be generated by the methods of Needleman and Wunsch(3), of Sellers(4), and the "local homology" method of Smith and Waterman(5). The Smith and Waterman alignment algorithm is also the most sensitive method available for identifying similarities between weakly related sequences.

Mapping and Searching

Mapping is available in several formats. Graphic maps display all of the cuts for each restriction enzyme on parallel lines. This graphic map facilitates selection of enzymes for isolating any region of a sequenced DNA molecule. Sorted maps in tabular format arrange the fragments from any digestion in order of molecular weight to show which fragments are similar in size and thus likely to be confused in gels. Another frequently used mapping format, designed by Frederick Blattner(6), displays the enzyme cuts above the original DNA sequence. Both strands of the DNA and all six frames of translation are shown.

All mapping programs will search for user-specified sequences, allowing features to be marked at the appropriate position on a restriction map. The mapping and searching programs can be used to aid site-specific mutagenesis experiments by showing where mutations could generate new restriction sites. All of the positions in a sequence where a synthetic probe could pair with one or more mismatches can also be located. Sequences related to less precisely defined features such as promoters or intervening sequence splice sites, can be located with a program that uses a consensus sequence as a probe. The

Table 1

Programs Available from UWCCG

Name	Function
DotPlot ⁺	makes a dot plot by method of Maizel and Lenk(2)
Cap	finds optimal alignment by method of Needleman and Wunsch(3)
BestFit	finds optimal alignment by method of Smith and Waterman(5)
MapPlot ⁺	shows restriction map for each enzyme graphically
MapSort	tabulates maps sorted by fragment position and size
Map	displays restriction sites and protein translations above and below the original sequence(Blattner,6)
Consensus	creates a consensus table from pre-aligned sequences
FitConsensus	finds sequences similar to a consensus sequence using a consensus table as a probe
Find	finds sites specified interactively
Stemloop	finds all possible stems (inverted repeats) and loops
Fold*	finds an RNA secondary structure of minimum free energy by the method of Zuker(7)
CodonPreference ⁺	plots the similarity between the codon choices in each reading frame and a codon frequency table(8)
CodonFrequency	tabulates codon frequencies
Correspond	finds similar patterns of codon choice by comparing codon frequency tables (Grantham et al,9)
TestCode ⁺	finds possible coding regions by plotting the "TestCode" statistic of Fickett(10)
Frame ⁺	plots rare codons and open reading frames(8)
PlotStatistics ⁺	plots asymmetries of composition for one strand
Composition	measures composition, di and trinucleotide frequencies
Repeat	finds repeats (direct, not inverted)
Fingerprint	shows the labelled fragments expected for an RNA fingerprint
Seqed	screen oriented sequence editor for entering, editing and checking sequences
Assemble	joins sequences together
Shuffle	randomizes a sequence maintaining composition
Reverse	reverses and/or complements a sequence
Reformat	converts a sequence file from one format to another
Translate	translates a nucleotide into a peptide sequence
BackTranslate	translates a peptide into a nucleotide sequence
Spew	sends a sequence to another computer
GetSeq	accepts a sequence from another computer
Crypt	encrypts a file for access only by password
Simplify	substitutes one of six chemically similar amino acid families for each residue in a peptide sequence
Publish	arranges sequences for publication
Poster ⁺	plots text (for labelling figures and posters)
OverPrint	prints darkened text for figures with a daisy wheel printer

⁺ requires a Hewlett Packard Series 7221 terminal plotter

* Fold is distributed by Dr. Michael Zuker not UWCCG.

Nucleic Acids Research

mapping programs can also be used on protein sequences to identify the peptides resulting from proteolytic cleavage.

Secondary Structure

Three programs are available to examine secondary structure in nucleic acids. The program StemLoop identifies all inverted repeats. An implementation of Dr. Michael Zuker's Fold program(7) finds an RNA secondary structure of minimum free energy based on published values of stacking and loop destabilizing energies. The "dot plot" comparison (mentioned above) of a sequence compared to its opposite strand gives a graphic picture of the pattern of inverted repeats in a sequence.

Analysis of Composition and the Location of Genetic Domains

Regions of a sequence with non-random base distribution can be displayed with three graphic tools designed to identify genetic domains. The program CodonPreference(8) identifies potential coding regions by searching through each reading frame for a pattern of preferred codon choices. The CodonPreference plot predicts the level of translational expression of mRNAs and helps identify frame shifts in DNA sequence data. Patterns of codon choice can be compared with the program Correspond(9). When a strong pattern of codon preferences is not expected, the "TestCode" statistic of Fickett(10) can be plotted to show regions of compositional constraint at every third base. Another program plots asymmetries of composition by strand. Strand asymmetries have been associated with genetic domains by several authors(11)(12). A fourth program called Frame marks the positions of rare codons and open reading frames on a graph showing all six reading frames.

Several tools are available to measure content and to count dinucleotide, trinucleotide, neighbor and repeat frequencies. A program that predicts RNA fingerprint patterns and another that tabulates codon frequencies complete the group of programs that analyze composition.

Sequence Manipulation

Sequences may be entered, assembled, edited, reversed, randomized, reformatted, translated, back-translated, documented, transferred, or encrypted rapidly with a large set of sequence manipulation tools.

A screen-oriented editor is available that allows sequences to be entered and checked. After a sequence is entered, it may be reentered for proofreading. Whenever a reentered base is at variance with the original, the terminal bell rings and the position is marked. Existing sequences can be edited quickly by moving directly to a sequence position specified by either a coordinate or a sequence pattern. The program can reassign the terminal's

keys to place G, A, T and C conveniently under the fingers of one hand in the same order as the lanes of a sequencing gel.

Programs are available for changing sequence file format. Sequence data from any source can be used in UWGCC programs, and sequence files maintained with UWGCC software can be converted for use in other non-UWGCC programs. For instance, the programs of Roger Staden(13) or Intelligenetics Inc.(14) could be used to assemble a sequence from the sequences of many small sub-fragments generated by DNAase I digestion. The assembled sequence could then be reformatted for use in any UWGCC program. A program is available that transfers sequences to and from other computers.

Sequence Publication

A program, Publish, will format sequences into figures. Publish has alternatives for line size, numbering, scaling, translation and comparison to other sequences. Poster is a program that will plot text on figures.

GENERAL FEATURES OF UWGCC SOFTWARE

Interactive Style

Each program is run by simply typing its name. Every parameter required by the program is obtained interactively. Questions are answered with a file name, a yes, a no, a number, or a letter from a menu. Default answers are displayed. Programs are insensitive to absurd answers and will ask the question again if, for instance, you name a file that does not exist or if you use a nonnumeric character when typing a number. Special features such as plotting features oriented to publication, are obtained by using an extra word next to the program's name when the program is run. Thus parameter queries are kept to a minimum for the normal use of each program.

Data

Both the NIH-GenBank(15) and the EMBL(16) nucleotide sequence data libraries are available "on-line" to any UWGCC program. A Search utility will locate sequences in the libraries by key word. A Find utility will locate library entries containing any specified sequence. A program is available that installs the new data sent periodically from GenBank and EMBL to update their data libraries.

All of the data in the system are stored in text files that can be read and modified easily. Every data file has an English heading describing the contents. The data files may be copied by each user for analysis or modification. Programs recognize and read user-modified input data automatically. Data files can be modified with any text editor.

Nucleic Acids Research

Sequence File Structure

Sequences are maintained in files that allow documentation and numbering both above and within the sequence. This file format is compatible with both of the nucleic acid sequence libraries and has been adopted as the standard sequence file format by the data base project at the European Molecular Biology Lab. Because genetic manipulations commonly involve linking several molecules of known sequence, UWCCG sequence files are designed to support concatenation by allowing comments to appear within the sequences at any location. Coding sequences or the boundaries between cloning vector and insert, for instance, can be marked within the sequence itself for immediate identification.

Sequence Symbols

All possible nucleotide ambiguities and all standard one-letter amino acid codes are part of the UWCCG symbol set that includes all alphabetic characters plus five additional characters. The proposed IUB-IUPAC standard nucleotide ambiguity symbols(17) are used for the mapping, searching and comparison programs. Lower case characters are used in sequences to indicate uncertainty as distinct from ambiguity. This allows the entire lexicon of symbols to be reused with same meaning, but with the prefix "maybe-." This reuse of the symbol set in lower case makes the uncertainty symbols more complete, understandable and visible.

Symbol Comparison

Sequence analysis programs generally make comparisons between sequence symbols (bases or amino acids) in order to find enzyme sites, create alignments, locate inverted repeats etc. These symbol comparisons are handled in several ways.

Symbol comparisons for alignment, comparison and secondary structure analysis are made by looking up a value in a symbol comparison table for the quality of the match. The table might contain 1's for matches and 0's for mismatches. If amino acids are being compared, however, a real number could be assigned at each position based on some previously assigned chemical similarity of the pair of residues or on the mutational distance between their codons. Standard symbol tables are provided by UWCCG, but the system is designed to allow each user to specify his own values.

Symbols comparisons for mapping and searching operations in nucleic acids are made by converting the IUB-IUPAC symbols into a binary code. The bits of this code represent G, A, T and C with ambiguity symbols causing more than one

bit to be set. A group of library functions identify overlap between the bits for each IUB-IUPAC symbol.

Documentation

Documentation is available both in printed form and on the terminal screen. A 350 page manual describes the operation of each program in detail, gives practical considerations and shows what will appear on the screen during a session with the program. Output files and plots are shown for the session. The data for the session shown in the documentation are included with the system so that the each program's operation can be checked. The "on-line" documentation is the same as the manual, but can be changed immediately when a program is modified.

All programs write output to files that are completely documented and sensibly organized for input to other programs. The input data, the program and the parameters used are clearly identified in every output file.

Procedure Library

UWGCG programs are written largely as calls to a library of 250 procedures designed to manipulate biological sequences. These procedures use data and file structures which have been designed to simplify program modification. For instance, standard operations such as reading sequences from files are always handled by a single library procedure. Thus a change in sequence file format requires only one subroutine to be modified for the new format to be acceptable to all of the programs in the system. Command procedures are available to help modify the library. The procedure library can be used by programs written in any language.

DISTRIBUTION OF UWGCG SOFTWARE

Intent

The intent of UWGCG is to make its software available at the lowest possible cost to as many scientists as possible.

Fees

A fee of \$2,000 for non-profit institutions or \$4,000 for industries is being charged for a tape and documentation for each computer on which UWGCG software is installed. While no continuing fee is required, UWGCG software, like the field it supports, is changing very rapidly. A consortium of industries and academic laboratories is planned to support the project in the future. The consortium will entitle its members to periodic updates and to influence the direction of new programming undertaken by UWGCG in return for a pledge of continuing financial support.

Nucleic Acids Research

Copyrights

UWCCG retains the copyrights to all of its software and UWCCG must be contacted before all or any part of the its software package is copied or transferred to any machine. UWCCG is, however, mandated to provide research tools to help scientists working in the area of molecular genetics and we are glad to see our source codes become the basis of further programming efforts by other scientists. Copyright can be removed for any program modified by more than 25% of its original effort.

Tape Format

The UWCCG package is usually distributed in VAX/VMS "backup" format on a 9 track magnetic tape recorded at 1600 bits/inch. The system consists of about 1000 files using about 20,000 blocks at 512 bytes/block. The current versions of the GenBank and EMBL nucleotide sequence data bases are normally included which add another 3,000 files and require another 20,000 blocks.

Upon request UWCCG will make a card image tape of all of the Fortran 77 programs and procedures for reading on computers other than the VAX. The card image tape is usually provided at 1600 bits/inch with 80 characters/record and 10 records/block. Adaptation of UWCCG software to systems other than VAX/VMS may take considerable effort.

Equipment Required

UWCCG programs and command procedures will run on a Digital Equipment Corporation (DEC) VAX computer that is using version 3.0 or greater of the DEC VMS operating system. A tape drive is necessary; a floating point accelerator and a DEC Fortran compiler are helpful, but not required. All programs can be run from a DEC VT52 or VT100 terminal. Seven programs, as noted in table 1, require a Hewlett Packard 7221 terminal plotter wired in series with the terminal. Several utilities support a daisy wheel compatible printer attached to the terminal's pass-through port, however, all programs write output files suitable for printing on any standard device.

Inquiries

Inquiries may be sent to John Devereux at the Laboratory of Genetics, University of Wisconsin, Madison, WI, USA 53706, (608) 263-8970. UWCCG is not licensed to distribute Fold(7), but the UWCCG implementation is available from Michael Zuker, Division of Biological Sciences, National Research Council of Canada, 100 Sussex Drive, Ottawa, Canada, K1A 0R6 (613) 992-4182.

ACKNOWLEDGEMENTS

UWCCG was started with software written for Oliver Smithies' laboratory

with NIH support from grants GM 20069 and AM 20120. UWCCG is directed by John Devereux and is operated as a part of the Laboratory of Genetics with the advice of a steering committee consisting of Richard Burgess, James Dahlberg, Walter Fitch, Oliver Smithies and Millard Susman. UWCCG is currently supported with intramural funds and with fees paid by the faculty and industries using the facility in Madison. This article is paper number 2684 from the Laboratory of Genetics, University of Wisconsin.

*Current address: Silicon Graphics Inc., 630 Clyde Court, Mountain View, CA 94043, USA

REFERENCES

1. Kernighan, B.W. and Plauger, P.J. (1976) Software Tools, Addison-Wesley Publishing Company, Reading, Massachusetts.
2. Maizel, J.V. and Lenk, R.P. (1981) Proceedings of the National Academy of Sciences USA 78, 7665-7669.
3. Needleman, S.B. and Wunsch, C.D. (1970) Journal of Molecular Biology 48, 443-453.
4. Sellers, P.H. (1974) SIAM Journal on Applied Mathematics 26, 787-793.
5. Smith, T.F. and Waterman, M.S. (1981) Advances in Applied Mathematics 2, 482-489.
6. Schroeder, J.L. and Blattner, F.R. (1982) Nucleic Acids Research 10, 69-84, Figure 1.
7. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Research 9, 133-148.
8. Gribskov, M., Devereux, J. and Burgess, R.R. "The Codon Preference Plot: Graphic Analysis of Protein Coding Sequences and Gene Expression," submitted to Nucleic Acids Research.
9. Grantham, R. Gautier, C. Guoy, M. Jacobzone, M. and Mercier R. (1981) Nucleic Acids Research 9(1), r43-r74.
10. Fickett, J.W. (1982) Nucleic Acids Research 10, 5303-5318
11. Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L., and S. Shen, (1981) Cell 26, 345-353.
12. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucleic Acids Research 11, 2205-2220.
13. Staden, R. (1980) Nucleic Acids Research 8, 3673-3694.
14. Clayton, J. and Kedes, L. (1982) Nucleic Acids Research 10, 305-321.
15. The GenBank(TM) Genetic Sequence Data Bank is available from Wayne Rindone, Bolt Beranek and Newman Inc., 10 Moulton Street, Cambridge, Massachusetts 02238, USA.
16. The EMBL Nucleotide Sequence Data Library is available from Greg Hamm, European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, West Germany.
17. Personal communication from Dr. Richard Lathe, Transgene SA, 11 Rue Humann, 67000 Strasbourg, France.

Exhibit 13

Cloning and sequence analysis of rat hepsin, a cell surface serine proteinase

David Farley, Françoise Reymond and Hanspeter Nick

Pharmaceuticals Research, Ciba-Geigy Ltd., Basel (Switzerland)

(Received 11 February 1993)

Key words: Hepsin; Serine proteinase; Proteinase, membrane-bound; cDNA sequence; (Rat liver)

A cDNA coding for the rat serine proteinase hepsin was isolated and its nucleotide sequence has been determined. The cDNA was 1739 nucleotides long and contained an open reading frame encoding a protein consisting of 416 amino-acid residues. The deduced amino-acid sequence of the rat enzyme was very similar to the human hepsin sharing an amino-acid sequence identity of 88.7%. Hydropathy plots reveal the presence of a short hydrophobic region close to the N-terminus believed to be a transmembrane domain which anchors the proteinase on the cell surface. The predicted sequence contains the His, Asp and Ser residues which make up the catalytic triad common to all serine proteinases.

Hepsin is a membrane-bound serine proteinase which was originally identified from cDNA clones isolated from human liver libraries [1]. The role of this proteinase is not known and the protein is poorly characterized with respect to its physical characteristics and substrate specificity. Human hepsin deduced from the encoding cDNA consists of 417 amino-acid residues and contains a short hydrophobic region near the amino-terminus believed to be a membrane spanning region. Immunostaining studies of cultured HepG2 cells demonstrate that hepsin is localized on the outer cell membrane surface with its NH₂-terminal side facing the cytosol and the carboxyl or catalytic side at the cell surface [2,3]. In this paper we report the cloning and sequence of the rat liver hepsin gene and compare structural similarities with human hepsin and other serine proteinases.

A rat liver cDNA library (Stratagene, No. 936507) was screened with a labeled DNA probe corresponding to 137 nucleotides at the 3'-end of the rat hepsin cDNA. This cDNA probe had previously been isolated attached to a rat 5- α -reductase cDNA [4]. Six positive clones were isolated after screening about $4.5 \cdot 10^5$ phage plaques. Restriction analysis of the DNA from the positive plaques revealed that the largest

insert was almost 1800 nucleotides in length. This *Eco*RI fragment was then subcloned into the plasmid pBSK-(Stratagene). The DNA insert was self-ligated, fractionated by sonication, subcloned into M13mp18 and both strands were sequenced using the dideoxy chain termination method [5].

The nucleotide sequence and the deduced amino-acid sequence for rat hepsin are shown in Fig. 1. The cDNA presented here is 1739 nucleotides in length and contains 184 nucleotides of untranslated sequence at the 5'-end, an open reading frame consisting of 1248 nucleotides encoding a protein of 416 amino-acid residues, a TGA stop codon, 304 nucleotides at the 3'-end and 33 adenine residues believed to make up the poly(A) tail. Based on the cDNA sequence, rat hepsin would have a predicted molecular mass of 44 930 Da and contains one potential N-linked carbohydrate attachment site at Asn-111.

Alignment of the deduced amino-acid sequence of rat and human hepsin is shown in Fig. 2. The aligned amino-acids reveal a large degree of homology with about 89% of the amino-acid residues being identical. Rat hepsin is one amino-acid residue shorter at the amino-terminus than the human enzyme. Like human hepsin, rat hepsin contains a 27-amino-acid hydrophobic region which is characteristic of a transmembrane domain [6]. This region is believed to anchor the proteinase on the outer cell membrane in a specific orientation with the catalytic domain exposed to the extracellular environment. Hepsin does not possess an obvious signal sequence but does appear to be synthesized

Correspondence to: D. Farley, Ciba-Geigy, K125.117, 4002 Basel, Switzerland.

The nucleotide sequencing data reported in this paper will appear in the DDBJ, EMBL and GenBank Nucleotide Sequence Databases under the accession number X70900.

as an inactive precursor with an Arg-161-Ile-162 cleavage site involved in zymogen activation. Cleavage of this peptide bond results in a noncatalytic polypeptide consisting of 161 amino-acid residues and a carboxy-

terminal catalytic chain consisting of 255 residues that contains several highly conserved regions common to serine proteinases. By comparing the hepsin sequence presented here to other well-characterized serine pro-

1	GCAGGCCCA																						
11	CCCTGCTGGCTGCTGCTGCCACCTTCCCTCCGGGCTGCCCGCTGCTGTGGGACACCATGCCCTTCCAGGCCCGGAGACTAAC																						
98	CCCAAACTGCACCATCTCCGGGAACCCAGGGTTCGGCCCCAGCCCAACAGGTCAACCTGGGAATCATTAACAAGAGTCCCTGAC																						
	M	A	K	E	G	G	R	T	A	P	C	C	S	R	P	K	V	A	A	L	T	V	22
185	ATG	GCG	AAG	GAG	GGT	GGC	CGG	ACT	GCA	CCA	TGC	TGT	TCC	AGA	CCC	AAG	GTG	GCA	GCT	CTC	ACT	GTG	
	G	T	L	L	F	L	T	G	I	G	A	A	S	W	A	I	V	T	I	L	L	R	44
251	GGG	ACC	CTG	CTG	TTC	CTG	ACA	GGC	ATT	GGG	GCT	GCG	TCC	TGG	GCC	ATT	GTG	ACC	ATC	CTA	CTA	CGG	
	S	D	Q	E	P	L	Y	Q	V	Q	L	S	P	G	D	S	R	L	L	V	L	D	66
317	AGT	GAC	CAG	GAG	CCA	CTG	TAC	CAA	GTG	CAG	CTC	AGT	CCC	GGG	GAC	TCT	CGA	CTT	TTG	GTG	TTG	GAC	
	K	T	E	G	T	W	R	L	L	C	S	S	R	S	N	A	R	V	A	G	L	G	88
383	AAG	ACA	GAG	GGA	ACG	TGG	AGG	CTG	CTG	TGC	TCC	TCA	CGC	TCC	AAC	GCC	AGG	GTA	GCA	GGG	CTC	GGC	
	C	E	E	M	G	F	L	R	A	L	A	H	S	E	L	D	V	R	T	A	G	A	110
449	TCT	GAG	GAG	ATG	GGC	TTT	CTC	AGG	GCT	CTG	GCG	CAC	TCA	GAG	CTG	GAT	GTG	CGA	ACC	GCG	GGC	GCC	
	N	G	T	S	G	F	F	C	V	D	E	G	G	L	P	L	A	Q	R	L	L	D	132
515	AAC	GGC	ACA	TCG	GGC	TTC	TTC	TGC	GTG	GAC	GAG	GGC	GGT	CTG	CCT	CTG	GCT	CAG	CGG	TTG	CTG	GAT	
	V	I	S	V	C	D	C	P	R	G	R	F	L	T	A	T	C	Q	D	C	G	R	154
581	GTC	ATC	TCT	GTA	TGC	GAC	TGT	CCT	AGA	GGC	CGA	TTC	CTG	ACT	GCC	ACC	TGC	CAA	GAC	TGT	GGC	CGC	
	R	K	L	P	V	D	R	I	V	G	G	Q	D	S	S	L	G	R	W	P	W	O	176
647	AGG	AAG	CTG	CCG	GTG	GAT	CGC	ATT	GTG	GGG	GGC	CAG	GAC	AGC	AGC	CTG	GGA	AGA	TGG	CCA	TGG	CAG	
	V	S	L	R	Y	D	G	T	H	L	C	G	G	S	L	L	S	G	D	W	V	L	198
713	GTC	AGC	CTG	CGT	TAT	GAT	GGG	ACC	CAC	CTC	TGT	GGG	GGA	TCC	CTG	CTG	TCC	GGG	GAC	TGG	GTA	CTG	
	T	A	A	H	C	F	P	E	R	N	R	V	L	S	R	W	R	V	F	A	G	A	220
779	ACC	GCT	GCA	CAC	TGC	TTT	CCA	GAG	AGG	AAC	CGG	GTC	CTG	TCT	CGG	TGG	CGA	GTA	TTT	GCT	GGT	GCT	
	V	A	R	T	S	P	H	A	V	Q	L	G	V	Q	A	V	I	Y	H	G	G	Y	242
845	GTA	GCC	CGG	ACC	TCA	CCT	CAT	GCC	GTG	CAG	CTG	GGG	GTT	CAG	GCT	GTG	ATC	TAT	CAT	GGG	GGC	TAC	
	L	P	F	R	D	P	T	I	D	E	N	S	N	D	I	A	L	V	H	L	S	S	264
911	CTT	CCC	TTT	CGA	GAC	CCT	ACT	ATC	GAC	GAA	AAC	AGC	AAT	GAC	ATT	GCC	CTG	GTC	CAC	CTC	TCT	AGC	
	S	L	P	L	T	E	Y	I	O	P	V	C	L	P	A	A	G	Q	A	L	V	D	286
977	TCC	CTG	CCT	CTC	ACA	GAA	TAC	ATC	CAG	CCG	GTT	TGT	CTC	CCT	GCT	GCG	GGA	CAG	GCC	CTG	GTG	GAC	
	G	K	V	C	T	V	T	G	W	G	N	T	Q	F	Y	G	Q	Q	A	V	V	L	308
1043	GGC	AAG	GTC	TGT	ACA	GTG	ACC	GGC	TGG	GGT	AAC	ACA	CAG	TTC	TAT	GGC	CAG	CAA	GCT	GTG	GTG	CTC	
	Q	E	A	R	V	P	I	I	S	N	E	V	C	N	S	P	D	F	Y	G	N	O	330
1109	CAA	GAG	GCC	CGG	GTC	CCC	ATC	ATA	AGC	AAC	GAA	GTT	TGC	AAC	AGC	CCC	GAC	TTC	TAC	GGG	AAT	CAG	
	I	K	P	K	M	F	C	A	G	Y	P	E	G	G	I	D	A	C	Q	G	D	S	352
1175	ATC	AAA	CCC	AAG	ATG	TTC	TGT	GCT	GGC	TAT	CCT	GAG	GGT	GGT	ATT	GAT	GCA	TGC	CAG	GGT	GAC	AGC	
	G	G	H	F	V	C	E	D	R	I	S	G	T	S	R	W	R	L	C	G	I	V	374
1241	GGA	GGC	CAC	TTT	GTA	TGT	GAG	GAC	AGA	ATC	TCT	GGA	ACA	TCA	AGA	TGG	CGG	CTC	TGC	GGC	ATT	GTA	
	S	W	G	T	G	C	A	L	A	R	K	P	G	V	Y	T	K	V	I	D	F	R	396
1307	AGC	TGG	GGT	ACG	GGC	TGT	GCT	TTC	CCC	CGG	AAG	CCG	GGA	GTG	TAC	ACC	AAA	GTC	ATT	GAC	TTC	CGG	
	E	W	I	F	Q	A	I	K	T	H	S	E	A	T	G	M	V	T	Q	P	Stop		416
1373	GAG	TGG	ATC	TTC	CAG	GCC	ATA	AAG	ACT	CAC	TCC	GAA	GCT	ACC	GGC	ATG	GTA	ACT	CAG	CCC	TGA	CCC	
1439	GCCCTCATCGCCTGCTCCGGGCTGCTCCAGCATCCAGAGTCAGAGTTGGTCTCGTGCTCCAGCCGACGTCGGCAGGGCTCCACACTG																						
1526	GGCCTCACATGGAACGGTTTCTGCTCGGATCCAGTCCATAGATCCAAGGATGCTGGGTCCAAGGACCTCTCTTCCACAGTGGCCGG																						
1613	CCCACTCAATCCACGGGCAATGGCTCACCTCCCAACCCCATGTAAATATTACTCTGCTCTGGGGGCTGCTTTCGAGGGCCCC																						
1700	TTGTGCGGATGCTCTTTAAATAATAAAGGTGGTTTTGATT																						

Fig. 1. cDNA sequence and predicted amino-acid sequence of rat hepsin. Nucleotides are numbered at left and amino-acid residues at right. The predicted transmembrane domain is underlined and (▼) represents the proposed zymogen activation cleavage site. The catalytic residues are starred and a potential N-linked glycosylation site is indicated by (●).

Rat	MAKGGRTAPCCSRPKVAALTVGTLTGTGAASWAVTLLASDQEPYQVLS PGDS	60
Hum	MAQ.....V.....A.....L.....A.....AV.....P.....V.....SA.....A	61
Rat	ALLVLDKTDGTMRLICSSRSNARVAGLOCEDEGFLRALAHSELDVRTAGANGTSGFFCVD	120
Hum	..M..F.....S.....T.....	121
Rat	EOGLPLAQRLLDVI SVDC PRORPLTATCQDCGRKL PVDRTVGGQSSSLGRPMQVSLR	180
Hum	..R...HT...E.....A..I.....R..T.....	181
Rat	YDQTHLCGGSLSGDMVLTAAKCF PERNRVLSRRVFPAGAVARTSPHAYQLGVQAVIYKQ	240
Hum	...A.....QA...GL.....V.....	241
Rat	GYLPPRPDPTIDENSNDIALVHLSSSLPLTEYIQPVCLPAAGQALVDGKVCVTVMGMDTOP	300
HumNSE.....P.....I.....Y	301
Rat	YQQQAVVLQEARVPITISNEVCHSPDFYGNQIKPKQFCAGYPEGGIDACQGDSSGHPVCEID	360
HumG.....D.....GA.....P.....	361
Rat	RISGTSRWLCCGVSMCTGCALARKPGVYTKVIDPREWIPOAKTHSEATGMVTOP	416
Hum	S..R..P.....Q.....S.....L	417

Fig. 2. Comparison of the deduced amino-acid sequences of rat and human hepsin. Residues in the human sequence that are identical to those of the rat are represented by a single dot and differences are indicated.

teinases, one can predict that the two conserved cysteine residues at positions 152 and 276 are involved in a disulfide linkage between the noncatalytic and catalytic chains of hepsin. Many interesting similarities of hepsin to other serine proteinases have already been considered by Leytus et al. [1] in their description of human hepsin.

Proteinases are involved in many biological processes such as blood coagulation, fibrinolysis and complement activation [7]. However, the biological role of hepsin remains unclear since its enzymatic specificity

and physiological substrates are presently unknown. Analysis of the amino-acid sequence of hepsin reveals several key residues which are similar to trypsin especially in the highly conserved sequences which surround the catalytic site. Although substrate specificity is unknown, the presence of an Asp at position 346 would suggest that hepsin exhibits trypsin like activity since a similar residue is found in trypsin at the bottom of the substrate binding pocket [8]. The precise role of this enzyme will remain a subject of speculation until the native enzyme can be purified and further characterized.

References

- 1 Leytus, S.P., Loeb, K.R., Hagen, F.S., Kurachi, K. and Davie, E.W. (1988) *Biochemistry* 27, 1067-1074.
- 2 Tsuji, A., Torres-Rosado, A., Arai, T., LeBeau, M.M., Lemons, R.S., Chou, S.H. and Kurachi, K. (1991) *J. Biol. Chem.* 266, 16948-16953.
- 3 Tsuji, A., Torres-Rosado, A., Arai, T., Chou, S.H. and Kurachi, K. (1991) *Biomed. Biochim. Acta* 50, 791-793.
- 4 Ordman, A., Farley, D., Meyhack, B. and Nick, H. (1991) *J. Steroid Biochem. Mol. Biol.* 39, 487-492.
- 5 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
- 6 Hartmann, E., Rapoport, T.A. and Lodish, H.F. (1989) *Proc. Natl. Acad. Sci. USA* 86, 5786-5790.
- 7 Neurath, H. (1986) *J. Cell Biochem.* 32, 35-49.
- 8 Stroud, R.M., Kay, L.M. and Dickerson, R.E. (1974) *J. Mol. Biol.* 83, 185-208.

Bio
C 1

BB.

A
ap
ele
rat
sugsuc
vol
be
ne
cle
of
mc
chl
ep
ph
ne
ex
lar
ch
tis
spicC
me
cC
pa
prCo
nal
ma
Th
sul
ac

Exhibit 14

Localization of the mosaic transmembrane serine protease corin to heart myocytes

John D. Hooper¹, Anthony L. Scarman¹, Belinda E. Clarke², John F. Normyle¹ and Toni M. Antalis¹

¹Cellular Oncology Laboratory, Queensland Institute of Medical Research, Brisbane, Queensland, Australia;

²Department of Anatomical Pathology, The Prince Charles Hospital, Chermside, Queensland, Australia

Corin cDNA encodes an unusual mosaic type II transmembrane serine protease, which possesses, in addition to a trypsin-like serine protease domain, two frizzled domains, eight low-density lipoprotein (LDL) receptor domains, a scavenger receptor domain, as well as an intracellular cytoplasmic domain. In *in vitro* experiments, recombinant human corin has recently been shown to activate pro-atrial natriuretic peptide (ANP), a cardiac hormone essential for the regulation of blood pressure. Here we report the first characterization of corin protein expression in heart tissue. We generated antibodies to two different peptides derived from unique regions of the corin polypeptide, which detected immunoreactive corin protein of approximately 125–135 kDa in lysates from human heart tissues. Immunostaining of sections of human heart showed corin expression was specifically localized to the cross striations of cardiac myocytes, with a pattern of expression consistent with an integral membrane localization. Corin was not detected in sections of skeletal or smooth muscle. Corin has been suggested to be a candidate gene for the rare congenital heart disease, total anomalous pulmonary venous return (TAPVR) as the corin gene colocalizes to the TAPVR locus on human chromosome 4. However examination of corin protein expression in TAPVR heart tissue did not show evidence of abnormal corin expression. The demonstrated corin protein expression by heart myocytes supports its proposed role as the pro-ANP convertase, and thus a potentially critical mediator of major cardiovascular diseases including hypertension and congestive heart failure.

Keywords: serine protease; corin; heart; pro-atrial natriuretic peptide (pro-ANP); TAPVR.

Serine proteases are found in all living organisms, ranging from viruses to humans [1], where they serve important and varied biological functions in situations requiring limited proteolysis. Their activities impact on areas as diverse as hemostasis, tissue remodelling and wound repair, inflammation, angiogenesis, fibrinogenesis and fibrinolysis. Cell surface serine proteases have been associated largely with extracellular matrix degradation, but there are emerging roles for these proteases in generating bioactive matrix protein fragments, influencing the release, the activation and bioavailability of growth factors and in shedding of cell surface proteins [2–6].

Many serine proteases are mosaic proteins comprising multiple, structurally distinct domains necessary for regulating enzymatic activity. Circulating serine proteases of the blood coagulation (e.g. prothrombin and factor X) [7], fibrinolysis (e.g. plasminogen activators) [8] and complement (e.g. C1r and C1s) [9] systems are well characterized examples of mosaic proteins. While the vast majority of known serine proteases are secreted, more recently some serine proteases have been found to possess integral transmembrane domains. The proteins enteropeptidase [10], hepsin [11] and most recently, TMPRSS2

[12] are examples of mosaic serine proteases with type II transmembrane domains. These enzymes are positioned on the plasma membrane via a membrane spanning domain close to the N-terminus. In addition to membrane spanning and protease domains, enteropeptidase also contains two low-density lipoprotein (LDL) receptor domains, a meprin-like domain, two C1r-like domains and a truncated scavenger receptor domain. An LDL receptor domain and a scavenger receptor domain have also been identified in TMPRSS2 [12]. The functions of these domains have not been determined.

Serine proteases play important roles in several aspects of heart physiology and cardiovascular disease [13]. The mast cell serine protease chymase is believed to be the major converter of angiotensin (ang)I to angII in human heart tissue [14]. The involvement of angII in normal cardiac function as well as in heart ailments such as hypertrophy, heart failure and ischaemic heart disease is indicated by the finding that inhibition of the angiotensin converting enzyme (ACE), leads to beneficial outcomes for sufferers of these diseases [15]. However, ACE inhibitors block only 10–20% of angI conversion in heart tissue whereas the remaining activity is blocked by serine protease inhibitors [16]. The fibrinolytic serine proteases tissue-type plasminogen activator (tPA) and urokinase-type plasminogen activator (uPA) are also thought to be involved in the progression of heart disease. uPA is present at significantly elevated levels in the atherosclerotic lesions responsible for myocardial infarction and failure [17]. The reduction in tPA from arteriolar smooth muscle cells is linked to the development of coronary artery disease in transplanted hearts [18].

Our own work and that of Yan *et al.* [19] has led to the recent cloning of a cDNA encoding a novel, multidomain type II transmembrane serine protease from human heart. The

Correspondence to T. M. Antalis, Queensland Institute of Medical Research, Post Office Royal Brisbane Hospital, Brisbane, 4029, Queensland, Australia. Fax: + 61 73362 0107, Tel.: + 61 73362 0312, E-mail: toniA@qimr.edu.au

Abbreviations: LDL, low-density lipoprotein; ANP, atrial natriuretic peptide; TAPVR, total anomalous pulmonary venous return; tPA, tissue-type plasminogen activator; uPA, urokinase-type plasminogen activator; ang, angiotensin; ACE, angiotensin converting enzyme.

(Received 24 July 2000, revised 12 September 2000, accepted 4 October 2000)

predicted protein, corin, comprises two frizzled domains, eight LDL receptor domains, a truncated scavenger receptor domain, in addition to the extracellular trypsin-like serine protease domain [19]. Recent expression of recombinant corin demonstrates that it possesses pro-atrial natriuretic peptide (ANP) convertase activity [20], and thus may play a critical role in the regulation of hypertension. *In situ* hybridization studies of mouse embryonic heart showed that corin mRNA was expressed as early as day 9.5 and maintained its expression through the adult animal [19]. The corin gene was mapped to human chromosome 4p12–13 [19], near the locus for the congenital heart disease, total anomalous pulmonary venous return (TAPVR). Here we present data describing for the first time native corin protein expression and localization in human heart.

MATERIALS AND METHODS

Identification of corin cDNA by homology cloning

Homology cloning was performed by RT-PCR using degenerate oligonucleotides corresponding to conserved regions of serine proteases [21–24]. Total RNA was isolated from S1a cells [25] following treatment with TNF α and cycloheximide for 4 h. RNA (5 μ g) was reverse transcribed at 42 °C using AMV reverse transcriptase (Promega, Madison, WI) in the presence of oligo dT_{12–18} (0.25 μ g μ L⁻¹) (Pharmacia Biotech, Sweden), 50 mM Tris/HCl, pH 8.3, 50 mM KCl, 10 mM MgCl₂, 10 mM dithiothreitol and 0.5 mM spermidine in a total volume of 20 μ L. PCR was performed using 1 μ L of the reverse transcriptase reaction mixture, 500 ng of each primer, 10 mM Tris HCl, pH 8.3, 50 mM KCl, 1.5 mM MgCl₂, 0.2 mM dNTPs and 1–2 units of Taq polymerase (Perkin Elmer). The primers were as follows. Forward, 5'-ACAGAATTCTGGGTIGTIACI-GCIGCICAYTG-3'; reverse, 5'-ACAGAATTCAXIGGICCI-CCI(C/G)(T/A)XTCICC-3'; where X = A or G, Y = C or T; I = inosine).

Cycling conditions: 2 cycles of 94 °C for 2.5 min, 35 °C for 2.5 min and 72 °C for 3 min, followed by 33 cycles of 94 °C for 2.5 min, 57 °C for 2.5 min and 72 °C for 3 min, with a final extension at 72 °C for 7 min. PCR products of approximately 450 bp were ligated into pGEM-T (Promega, Madison, WI, USA), cloned and analysed by DNA sequencing. A DNA fragment was identified which represented the partial corin sequence (nucleotides 334–748). The cDNA was extended 333 nucleotides towards the 5' end by screening a cDNA library using two rounds of PCR and the nested oligonucleotides ATC2P3 and ATC2P1 in combination with the vector specific primer T7. The 3' end was extended to nucleotide 976 by two rounds of PCR and the nested oligonucleotides ATC2P4 and ATC2P5 in combination with the vector specific primer T3. The primer sequences are given below.

ATC2P1: 5'-GCGTGTCTGCATGAACACTG-3'; ATC2P2: 5'-ATGCCAAGCACCACCTTTCCA-3'; ATC2P3: 5'-ATAGTC-CACCACTGCTCGAC-3'; ATC2P4: 5'-TTAAGCTGCAAGA-GGGAGAG-3'.

The DNA sequence of this cDNA has been deposited in the DDBJ/Genbank/EMBL database under accession no. AF113248.

Heart tissue specimens

Tissues from explanted hearts with terminal heart failure were either snap frozen in liquid nitrogen (for RNA and protein analyses) or processed for routine histological examination. Six

paraffin embedded blocks of human heart tissue were obtained from autopsy cases with acute myocardial infarction. These blocks included both viable and nonviable myocardium. Procedures were in accordance with guidelines established by the National Health and Medical Research Council of Australia, Ethics Approval number EC9876(II).

Northern and Poly(A)⁺ RNA dot blot analyses

Human multiple tissue northern blots (Clontech, Palo Alto, CA, USA) contained 2 μ g of poly(A)⁺ RNA per lane. The blots were hybridized with a ³²P-dCTP labeled *Eco*RI digested DNA fragment encoding corin cDNA in ExpressHyb (Clontech) solution at 65 °C and washed to a final stringency of 0.2 \times NaCl/Cit, 0.1% SDS at 65 °C. The blot was reprobed with β -actin as a measure of loading in each lane. For the mouse tissue blot, total RNA was purified from mouse tissues, separated by denaturing gel electrophoresis and transferred to Hybond-N nylon membranes as described [26]. The blot was hybridized with the radiolabelled human corin DNA probe under lower stringency conditions in ExpressHyb solution at 55 °C and washed to a final stringency of 1 \times NaCl/Cit, 0.1% SDS at 55 °C. The mouse tissue blot was stained with ethidium bromide to confirm RNA loading in each lane.

Production of affinity purified anti-peptide polyclonal antibodies

Rabbit polyclonal antibodies were generated against corin specific peptides derived from nonhomologous hydrophilic regions within the corin amino-acid sequence. Two peptides, each containing a cysteine residue incorporated at the C-terminus, were synthesized (Auspep, Parkville, Australia) and conjugated to keyhole limpet hemocyanin using μ -maleimidobenzoic acid *N*-hydroxysuccinimide ester. The peptides were: A1: IQEQE-KEPRWLTLHSNWE-C, A2: GHMGNKMPFKLQEGE-C. Rabbit antisera was peptide-affinity purified using SulfoLink coupling gel (Pierce, Rockville, IL). The specificity of each antibody was tested against the immunogenic peptide by ELISA.

Western blot analysis

Frozen heart tissue (100 mg) was homogenized in lysis-binding buffer (Dynabeads mRNA Direct kit, Dynal) and spun at 13000g for 2 min. The protein pellet was dissolved in reducing SDS-sample buffer for Western blot analysis. Proteins were separated by SDS/PAGE on 10% acrylamide gels and transferred electrophoretically to Hybond-P membranes (Amersham, Aylesbury, UK). Membranes were blocked with 5% nonfat skim milk powder in Tris/NaCl (10 mM Tris/HCl, pH 7.0, 150 mM NaCl), incubated with affinity purified anti-peptide antibody, then with horseradish peroxidase conjugated sheep anti-(rabbit Ig) secondary antibody, and visualized by enhanced chemiluminescence (Amersham, Aylesbury, UK).

Immunohistochemistry

Paraffin sections (5 μ m) of formalin-fixed human heart were deparaffinized, then rehydrated before antigen retrieval in boiling 10 mM citric acid buffer, pH 6. After cooling, endogenous peroxidase activity was inhibited by 10 min incubation in 1% hydrogen peroxide. Non-specific antibody binding was blocked by incubating the sections in 4% nonfat skim milk powder in NaCl/P_i for 15 min, followed by 10%

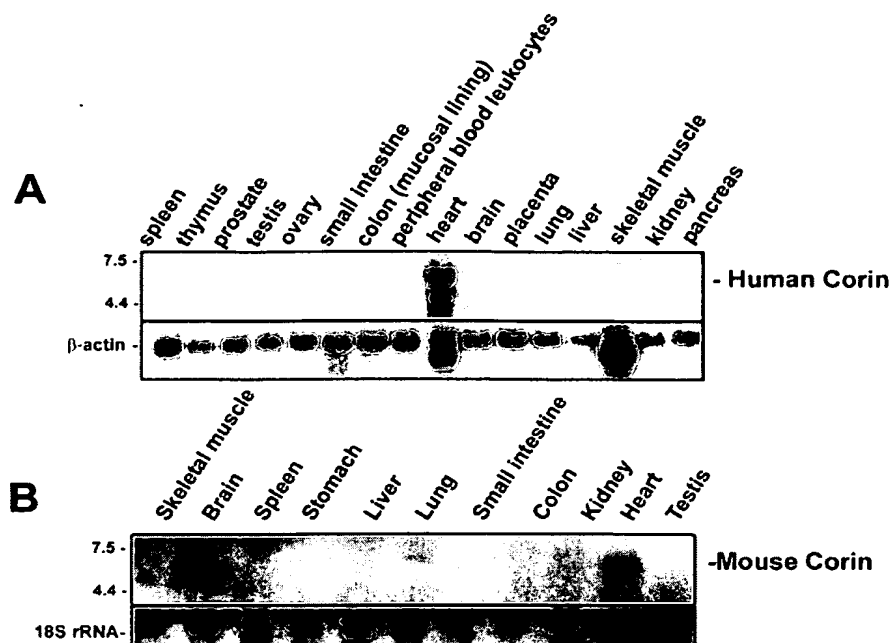


Fig. 1. Corin expression in human and mouse tissues. (A) Northern blot analysis of RNA isolated from a range of normal human tissues probed with 32 P-labelled corin cDNA. The levels of β -actin mRNA are shown as a control for loading. (B) Northern blot analysis of corin mRNA expression in a range of mouse tissues probed with 32 P-labelled human corin cDNA at reduced stringency. The levels of 18S ribosomal RNA are shown as a control for loading.

normal goat serum for 20 min. Affinity purified anticorin A1 (1 : 100; $150 \mu\text{g}\cdot\text{mL}^{-1}$) or A2 antibodies (1 : 50; $20 \mu\text{g}\cdot\text{mL}^{-1}$) were applied and incubated overnight in a humidified chamber at room temperature. Controls included sections incubated with no primary antibody or antibody that had been preadsorbed for 2 h at room temperature with $1 \mu\text{g}$ of the antigenic peptide. Following incubation with prediluted biotinylated goat anti-(rabbit Ig) Ig (Zymed, San Francisco, CA, USA), streptavidin–horseradish peroxidase (Zymed) was applied and color developed using the chromogen 3,3'-diaminobenzidine with hydrogen peroxide as substrate. The sections were counterstained in Mayers' haematoxylin.

RESULTS AND DISCUSSION

Isolation of human corin cDNA by homology cloning

A PCR-based homology cloning approach was employed to identify serine protease cDNAs expressed by the S1a cell line [25] which is resistant to tumor necrosis factor- α induced apoptosis. Degenerate primers designed to anneal to cDNA encoding the conserved regions surrounding the catalytic histidine and serine amino acids of serine proteases [21–23], were used to amplify and then clone a range of DNA fragments of approximately 450 bp. One clone, designated ATC2, was found to encode a novel serine protease. The cDNA was extended in the 5' and 3' directions by library screening and the DNA sequence was deposited in the DDBJ/Genbank/EMBL database (accession no. AF113248). This sequence was subsequently determined to be 100% identical to a recently reported cDNA encoding the serine protease, corin (accession no. AF133845) [19].

Corin mRNA is strongly expressed in heart

The tissue distribution of corin mRNA was examined by Northern blot analyses. Analysis of poly(A) $^{+}$ RNA from 16

normal human tissues showed a single transcript of approximately 5.1 kb detectable only in human heart (Fig. 1A). Examination of a range of mouse tissues also demonstrated specific expression of corin mRNA of approximately 5.1 kb only in mouse heart (Fig. 1B).

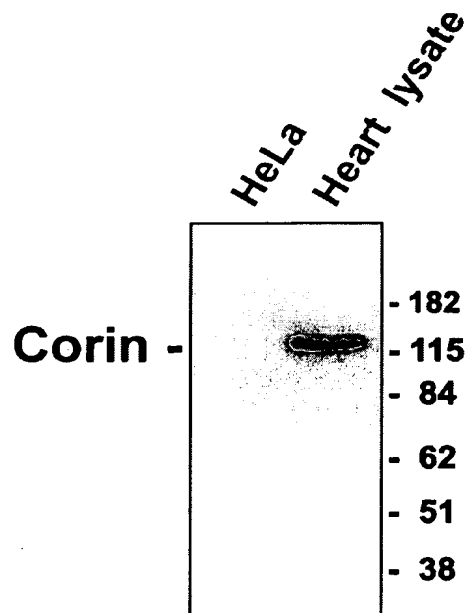


Fig. 2. Corin protein expression in human heart tissue by Western blot analysis. Immunoreactive corin protein of 125–135 kDa is detected in a protein lysate prepared from human heart tissue (Patient #7684), which is not detectable in a corin negative HeLa cell lysate. The blot was probed with anticorin antibody, AbA1, and visualized using enhanced chemiluminescence. The protein standards in kDa are as indicated.

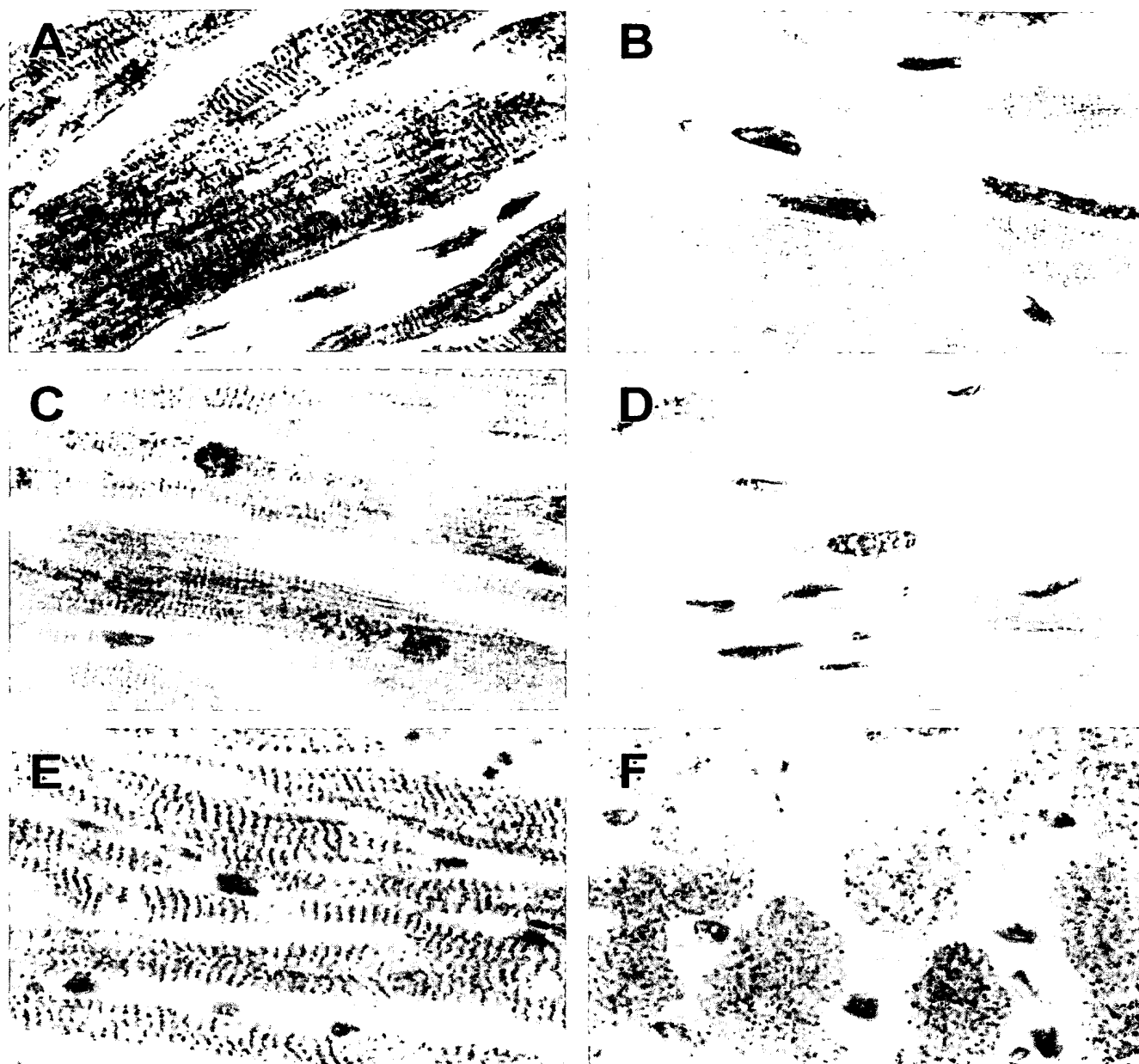


Fig. 3. Corin is localized to human heart myocytes by immunostaining. Immunohistochemical staining of human heart tissues was performed using the affinity purified anticorin peptide A1 or A2 polyclonal antibodies as primary antibodies. (A) a longitudinal section of a representative heart tissue from a transplant recipient (Patient #7684) stained with AbA1 showing intense staining in the cardiac myocytes; (B) as (A) except the primary antibody was preadsorbed with the immunogenic peptide, A1, for 2 h; (C) the same tissue as (A) except stained with the weaker staining antibody, AbA2. Apparent staining at the poles of the nuclei are deposits of the brown lipochrome pigment, lipofuscin. (D) the same tissue as (A–C) processed in the absence of primary antibody; (E) a longitudinal section of normal myocardium from a heart which contained an acute infarct elsewhere (Patient #A4–99R) stained with AbA1 showing intense staining corresponding to the cross striations; (F) staining of the same heart tissue as (E) with AbA1 showing intense staining in cross section. Photomicrographs (A–E) were taken at an original magnification of 100 \times .

Anti-corin antibodies detect corin in heart lysates

We generated polyclonal antibodies to two different peptides derived from unique regions of the corin polypeptide sequence in order to investigate its expression and localization in the heart. The first was a unique region within the serine protease catalytic domain between the conserved Asp and Ser

amino-acid residues (AbA1) and the second was contained within the scavenger receptor domain (AbA2). Immunoblot analysis of corin protein expression in human heart protein lysates showed a major immunoreactive band of 125–135 kDa (Fig. 2), which was not present in lysates from the negative control HeLa cell line. This molecular mass is slightly lower than that reported (\approx 150 kDa) for recombinant V5/His6

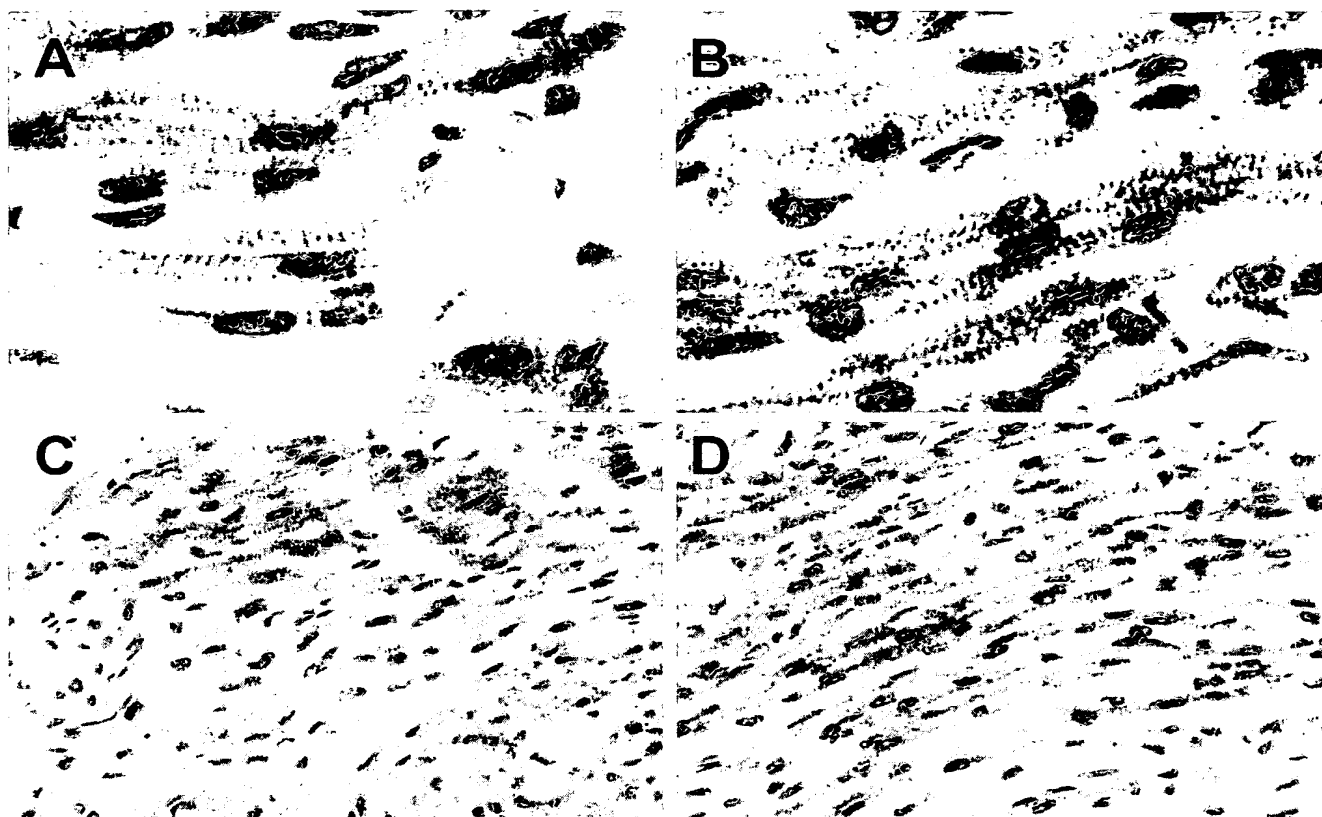


Fig. 4. Corin expression in neonate heart with TAPVR. Immunohistochemical staining of human neonate heart tissues was performed using the affinity purified anticorin peptide A1 polyclonal antibody as the primary antibody. (A) and (C) longitudinal sections of TAPVR heart tissue showing staining in the cardiac myocytes, corresponding to the cross striations; (B) and (D) longitudinal sections of a normal neonate heart showing a similar staining pattern in the cardiac myocytes. Photomicrographs (A) and (B) were taken at an original magnification of 100x and (C) and (D) were taken at an original magnification of 40x.

tagged corin expressed by human embryonic kidney 293 cells [20]. As the mature corin zymogen has a calculated mass of 116 kDa [19], it is likely that the mature corin polypeptide undergoes a post-translational processing event, possibly glycosylation. Consistent with this, there are 19 predicted N-linked glycosylation sites present in the extracellular domains of corin [19].

Corin is expressed by human heart myocytes

To investigate the localization of corin expression in human heart, immunohistochemical analyses were performed on human adult heart tissues. Corin was abundantly expressed in cardiac myocytes, with intense brown staining associated with cross striations seen in longitudinally sectioned myofibers (Fig. 3A). In some areas there was accentuation of the plasma membrane, consistent with an integral membrane localization of corin. This same pattern of staining was observed in sections taken from all areas of the myocardium. Control slides using the AbA1 polyclonal antibody in the presence of competing A1 peptide showed absence of this specific staining pattern (Fig. 3B). An identical, albeit weaker staining pattern was observed in experiments performed using the second corin-specific antibody (AbA2) (Fig. 3C). No staining was detected in the absence of antibody (Fig. 3D). Staining of a section of

viable myocardium from a heart containing an acute myocardial infarct showed a similar intense staining of the striations in cardiac myocytes (Fig. 3E) and a pinhead-like dot pattern when viewed in cross section (Fig. 3F). Necrotic heart tissue showed similar but much less intense staining (data not shown). Corin was not detected in sections of skeletal or smooth muscle (data not shown), suggesting that the function of corin is specifically related to cardiac muscle.

Corin protein expression in a patient with the congenital heart disease, TAPVR

The molecular mechanisms responsible for the developmental defect associated with the rare congenital heart disease TAPVR are not known. The location of the corin gene on human chromosome 4p12–13 [19] and the localization of the TAPVR locus to a 30 centimorgan interval on 4p13–q12 [26], suggested that corin may be a candidate for the TAPVR gene [19]. If corin plays a role in TAPVR, its expression may be lost or altered in TAPVR heart tissue. To explore this possibility, we examined corin protein expression in a TAPVR heart. The pattern of corin expression detected in this heart tissue (Fig. 4A,C) was similar to that observed in the adult heart and was identical to the pattern of corin staining in an age-matched neonate control heart (Fig. 4B,D). While this data is not consistent with a role

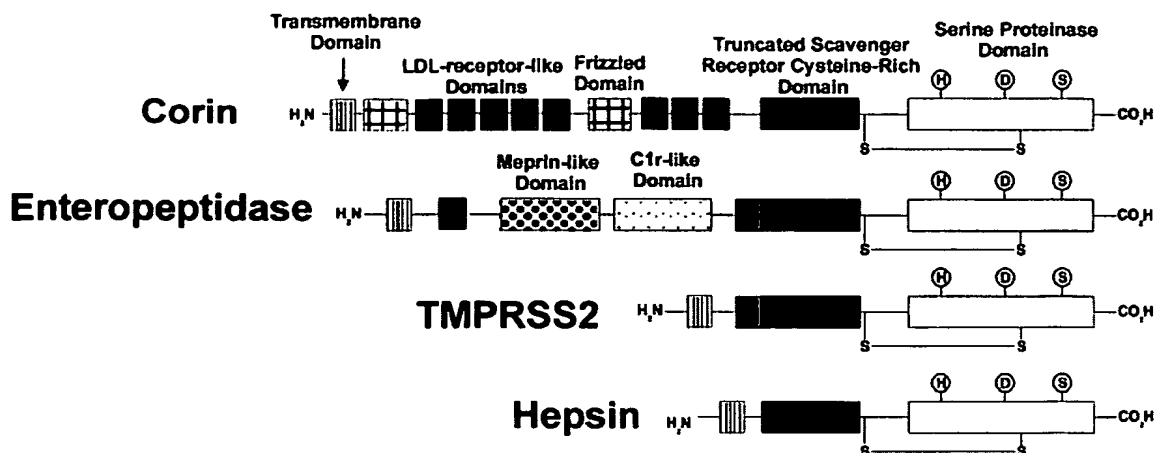


Fig. 5. Diagram showing domain structures of corin compared with other mosaic integral membrane proteins. The domains are as indicated. The catalytic serine protease residues are circled. The disulfide bond linking catalytic and pro-regions are marked.

for corin in TAPVR, it does not exclude the possibility that TAPVR is associated with more subtle alterations to the corin gene; for example point mutations, that would not be detected by this method.

Corin homology to other type II transmembrane proteases

As illustrated in Fig. 5, corin is a mosaic integral membrane protein possessing discrete domains. The intracellular, cytoplasmic domain contains two potential protein kinase C phosphorylation sites which may represent mechanisms for signal relay to or from the cell surface. Corin contains two frizzled domains. These domains function in other molecules as receptors for Wnt proteins, which are implicated in signal transduction during development [28]. Corin possesses eight LDL receptor domains which can mediate uptake of LDLs [29] and have also been shown to be involved in binding and internalization of protease/inhibitor complexes [30]. LDLs regulate the transport of cholesterol and play a major role in the development of heart disease. Corin possesses a scavenger receptor domain, which in other proteins, binds polyanionic molecules including modified lipoproteins, cell surface lipids and some sulfated polysaccharides [31]. The trypsin-like serine protease domain is located at the C-terminus.

Corin bears similarity to other known members of the integral membrane serine proteases as illustrated in Fig. 5. The corin serine protease domain is highly homologous to a multidomain integral-membrane serine protease found in the brush border of the intestine, enteropeptidase [32]. Enteropeptidase functions to activate digestive pancreatic enzymes released from the intestine. Activation of this cascade is critical, as illustrated by the life-threatening intestinal malabsorption that accompanies congenital deficiency of enteropeptidase [32]. Other proteases with homology to the corin serine protease domain are the integral-membrane serine proteases, TMPRSS2 and hepsin. Hepsin is a hepatic serine protease that has been demonstrated to activate Factor VII in the extrinsic blood coagulation pathway leading to thrombin formation, and has further been shown to be required for mammalian cell growth [33].

In summary, we have confirmed heart as a site of abundant corin mRNA expression and demonstrated for the first time the expression of corin as a 125–135 kDa protein in this tissue. In

addition, in heart we have localized corin protein to myocytes; the same cardiac cells expressing pro-ANP. These data support recently reported *in vitro* evidence that the corin proteolytic domain is the pro-ANP convertase [20] and thus, the proposal that corin has a role in regulating blood pressure. Possible additional functions of the serine protease domain and the functions of the other corin domains are not yet known. The putative phosphorylation sites in the cytoplasmic domain of corin may indicate that the intracellular domain of corin will be a target for phosphorylation and therefore may mediate signalling events from the cell surface. A better understanding of the role of corin in heart will provide insight into basic molecular mechanisms of cardiac function and could provide a rational target for both diagnostic and therapeutic applications.

ACKNOWLEDGEMENTS

This work was supported by grants from the Queensland Cancer Fund, Brisbane, Australia and the National Health and Medical Research Council of Australia. J. D. H. was supported by a John Earnshaw Scholarship from the Queensland Cancer Fund and by the Bancroft Scholarship, Queensland Institute of Medical Research.

REFERENCES

1. Rawlings, N.D. & Barrett, A.J. (1994) Families of serine peptidases. *Methods Enzymol.* **244**, 19–61.
2. Murphy, G. & Gavrilovic, J. (1999) Proteolysis and cell migration: creating a path? *Curr. Opin. Cell Biol.* **11**, 614–621.
3. LeMosy, E.K., Hong, C.C. & Hashimoto, C. (1999) Signal transduction by a protease cascade. *Trends Cell Biol.* **9**, 102–107.
4. Rifkin, D.B., Mazzieri, R., Munger, J.S., Noguera, I. & Sung, J. (1999) Proteolytic control of growth factor availability. *Acta Path. Microbiol. Immunol. Scand.* **107**, 80–85.
5. Dery, O. & Bunnett, N.W. (1999) Proteinase-activated receptors: a growing family of heptahelical receptors for thrombin, and trypsin. *Biochem. Soc. Trans.* **27**, 246–254.
6. Noel, A., Gilles, C., Bajou, K., Devy, L., Kebers, F., Lewalle, J.M., Maquoi, E., Munaut, C., Remacle, A. & Foidart, J.M. (1997) Emerging roles for proteinases in cancer. *Invasion Metastasis* **17**, 221–239.
7. Ichinose, A. & Davie, E.W. (1994) The Blood Coagulation Factors: Their cDNAs, Genes, and Expression. In *Hemostasis and Thrombosis: Basic Principles and Clinical Practice* (Colman, R.W.,

- Hirsh, J., Marder V.J. & Salzman, E.W., eds), pp. 19–54. J.B. Lippincott Company, Philadelphia, PA, USA.
8. Francis, C.W. & Marder, V.J. (1994) Physiologic Regulation and Pathologic Disorders of Fibrinolysis. In *Hemostasis and Thrombosis: Basic Principles and Clinical Practice* (Colman, R.W., Hirsh, J., Marder V.J. & Salzman, E.W., eds), pp. 1076–1103. J.B. Lippincott Company, Philadelphia, PA, USA.
9. Arlaud, G.J. & Thielens, N.M. (1993) Human complement serine proteases C1r and C1s and their proenzymes. *Methods Enzymol.* 223, 61–82.
10. Kitamoto, Y., Veile, R.A., Donis-Keller, H. & Sadler, J.E. (1995) Human complement serine proteases C1r and C1s and their proenzymes. *Biochemistry* 34, 4562–4568.
11. Tsuji, A., Torres-Rosado, A., Arai, T., Le Beau, M.M., Lemons, R.S., Chou, S.H. & Kurachi, K. (1991) Hepsin, a cell membrane-associated protease. Characterization, tissue distribution, and gene localization. *J. Biol. Chem.* 266, 16948–16953.
12. Paoloni-Giacobino, A., Chen, H., Peitsch, M.C., Rossier, C. & Antonarakis, S.E. (1997) Cloning of the TMPRSS2 gene, which encodes a novel serine protease with transmembrane, LDLRA, and SRCR domains and maps to 21q22.3. *Genomics* 44, 309–320.
13. Schussheim, A.E. & Fuster, V. (1997) Thrombosis, antithrombotic agents, and the antithrombotic approach in cardiac disease. *Prog. Cardiovascular Diseases* 40, 205–238.
14. Balcells, E., Meng, Q.C., Johnson, W.H. Jr, Oparil, S. & Dell'Italia, L.J. (1997) Angiotensin II formation from ACE and chymase in human and animal hearts: methods and species considerations. *Am. J. Physiol.* 273, H1769–H1774.
15. Wolny, A., Clozel, J.P., Rein, J., Mory, P., Vogt, P., Turino, M., Kiowski, W. & Fischli, W. (1997) Functional and biochemical analysis of angiotensin ii-forming pathways in the human heart. *Circ. Res.* 80, 219–227.
16. Bumpus, F.M. (1991) Angiotensin I and II. Some early observations made at the Cleveland Clinic Foundation and recent discoveries relative to angiotensin II Formation in human heart. *Hypertension* 18, 122–125.
17. Kienast, J., Padro, T., Steins, M., Li, C.X., Schmid, K.W., Hammel, D., Scheld, H.H. & Van De Loo, J.C. (1998) Relation of urokinase-type plasminogen activator expression to presence and severity of atherosclerotic lesions in human coronary arteries. *Thromb. Haemost.* 79, 579–586.
18. Labarrere, C.A., Pitts, D., Nelson, D.R. & Faulk, W.P. (1995) Vascular tissue plasminogen activator and the development of coronary artery disease in heart-transplant recipients. *N. Engl. J. Med.* 333, 1111–1116.
19. Yan, W., Sheng, N., Seto, M., Morser, J. & Wu, Q. (1999) Corin, a mosaic transmembrane serine protease encoded by a novel cDNA from human heart. *J. Biol. Chem.* 274, 14926–14935.
20. Yan, W., Wu, F., Morser, J. & Wu, Q. (2000) Corin, a transmembrane cardiac serine protease, acts as a pro-atrial natriuretic peptide-converting enzyme. *Proc. Natl Acad. Sci. USA* 97, 8525–8529.
21. Sakanari, J.A., Staunton, C.E., Eakin, A.E., Craik, C.S. & McKerrow, J.H. (1989) Serine proteases from nematode and protozoan parasites: isolation of sequence homologs using generic molecular probes. *Proc. Natl Acad. Sci. USA* 86, 4863–4867.
22. Elvin, C.M., Whan, V. & Riddles, P.W. (1993) A family of serine protease genes expressed in adult buffalo fly (*Haematobia irritans exigua*). *Mol. Gen. Genet.* 240, 132–139.
23. Elvin, C.M., Vuocolo, T., Smith, W.J., Eisemann, C.H. & Riddles, P.W. (1994) An estimate of the number of serine protease genes expressed in sheep blowfly larvae (*Lucilia cuprina*). *Insect Mol. Biol.* 3, 105–115.
24. Hooper, J.D., Nicol, D.L., Dickinson, J.L., Eyre, H.J., Scarman, A.L., Normyle, J.F., Stuttgen, M.A., Douglas, M., Loveland, K.A.L., Sutherland, G.R. & Antalis, T.M. (1999) Testisin, a new human serine proteinase expressed by premeiotic testicular germ cells and lost in testicular germ cell tumors. *Cancer Res.* 59, 3199–3205.
25. Dickinson, J.L., Bates, E.J., Ferrante, A. & Antalis, T.M. (1995) Plasminogen activator inhibitor type 2 inhibits tumor necrosis factor alpha induced apoptosis. Evidence for an alternate biological function. *J. Biol. Chem.* 270, 27894–27904.
26. Antalis, T.M. & Dickinson, J.L. (1992) Control of plasminogen activator inhibitor type 2 gene expression in the differentiation of monocytic cells. *Eur. J. Biochem.* 205, 203–209.
27. Bleyl, S., Nelson, I., Odelbury, S.J., Ruttenberg, H.D., Otterud, B., Leppert, M. & Ward, K. (1995) A gene for familial total anomalous pulmonary venous return maps to chromosome 4p13-q12. *Am. J. Hum. Genetics* 56, 408–415.
28. Cadigan, K.M. & Nusse, R. (1997) Wnt signaling: a common theme in animal development. *Genes Dev.* 11, 3286–3305.
29. Bujo, H., Yamamoto, T., Hayashi, K., Hermann, M., Nimpf, J. & Schneider, W.J. (1995) Mutant oocyte low density lipoprotein receptor gene family member causes atherosclerosis and female sterility. *Proc. Natl Acad. Sci. USA* 92, 9905–9909.
30. Kounnas, M.Z., Church, F.C., Argraves, W.S. & Strickland, D.K. (1996) Cellular internalization and degradation of antithrombin III-thrombin, heparin cofactor II-thrombin, and α 1-antitrypsin-trypsin complexes is mediated by the low density lipoprotein receptor-related protein. *J. Biol. Chem.* 271, 6523–6529.
31. Resnick, D., Chatterton, J.E., Schwartz, K., Slayter, H. & Krieger, M. (1996) Structures of class A macrophage scavenger receptors. Electron microscopic study of flexible, multidomain, fibrous proteins and determination of the disulfide bond pattern of the scavenger receptor cysteine-rich domain. *J. Biol. Chem.* 271, 26924–26930.
32. Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D.W. & Sadler, J.E. (1994) Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains. *Proc. Natl Acad. Sci. USA* 91, 7588–7592.
33. Torres-Rosado, A., O'Shea, K.S., Tsuji, A., Chou, S.H. & Kurachi, K. (1993) Hepsin, a putative cell-surface serine protease, is required for mammalian cell growth. *Proc. Natl Acad. Sci. USA* 90, 7181–7185.

Exhibit 15

Type II Transmembrane Serine Proteases

INSIGHTS INTO AN EMERGING CLASS OF CELL SURFACE PROTEOLYTIC ENZYMES*

Published, JBC Papers in Press, November 1, 2000,
DOI 10.1074/jbc.R000020200

John D. Hoopert, Judith A. Clements†, James P. Quigley§, and Toni M. Antalis||

From the ‡Centre for Molecular Biotechnology, Queensland University of Technology, Gardens Point, Brisbane 4000, Australia, §The Scripps Research Institute, La Jolla, California 92037, and the ||Cellular Oncology Laboratory, University of Queensland and the Queensland Institute of Medical Research, Brisbane 4029, Australia

Cell surface proteolysis has emerged as an important mechanism for the generation of biologically active proteins that mediate a diverse range of cellular functions. The proteolytic activities of membrane-anchored proteins, such as ADAMs¹ (1) and MT-MMPs (2), are thought to play central roles in cell surface-activating events. In contrast, most of the members of the serine protease family, one of the oldest characterized and largest multigene proteolytic families, are either secreted enzymes or sequestered in cytoplasmic storage organelles awaiting signal-regulated release. These serine proteases have well characterized roles in diverse cellular activities, including blood coagulation, wound healing, digestion, and immune responses, as well as tumor invasion and metastasis. However, during the last few years there has been an explosion in the identification of transmembrane proteins containing C-terminal extracellular serine protease domains. These enzymes are ideally positioned to interact with other proteins on the cell surface as well as soluble proteins, matrix components, and proteins on adjacent cells. In addition, these membrane-spanning proteases have cytoplasmic N-terminal domains, suggesting possible functions in intracellular signal transduction. This review delineates for the first time this emerging class of cell surface proteolytic enzymes, the type II transmembrane serine proteases (TTSPs), to highlight their structural features, expression profiles, and possible roles in mediating cell surface proteolytic events.

Structural Features of TTSPs

In mammals the TTSPs currently consist of 17 members (Table 1), of which seven are found in man. Enteropeptidase (also known as enterokinase) (3), because of its essential role in the processing of digestive proteases, was the first member of this group to be discovered nearly a century ago. The other more recently identified members include hepsin (4), human airway trypsin-like protease (HAT) (5), corin (6), MT-SP1 (7) (also known as matriptase (8)),

TMPRSS2 (9), and most recently TMPRSS4² (10). The only non-mammalian TTSP identified to date is the *Drosophila* protease stubble-stubloid (st-sb) (11). Mammalian orthologues have been reported for enteropeptidase (mouse (12), rat (13), cow (14), and pig (15)), hepsin (mouse (16) and rat (17)), corin (mouse, also known as LRP4 (18)), MT-SP1 (mouse, also known as epithin (19)), and TMPRSS2 (mouse, also known as epitheliasin (20)) (Table 1). The TTSPs share a number of common structural features including (i) a proteolytic domain, (ii) a transmembrane domain, (iii) a short cytoplasmic domain, and (iv) a variable length stem region containing modular structural domains, which links the transmembrane and catalytic domains (Fig. 1). It is this unique combination of domains that suggests novel roles for the TTSPs at the cell surface.

Proteolytic Domains—As is the case for the wider family of enzymes of the chymotrypsin (S1) fold,³ the proteolytic domains of the TTSPs share a high degree of amino acid sequence identity. In particular, the histidine, aspartate, and serine residues necessary for catalytic activity are present in highly conserved motifs. TTSPs are synthesized as single chain zymogens and are likely activated by cleavage following an arginine or lysine present in a highly conserved activation motif. Based on the predicted presence of a conserved disulfide bond linking the pro- and catalytic domains (Fig. 1), the TTSPs are likely to remain membrane-bound following activation. However, the isolation of soluble forms of enteropeptidase (21, 22), HAT (23), and MT-SP1 (24) suggests that the extracellular domains of at least some of the TTSPs may also be shed from the cell surface. Other cysteine residues conserved among the TTSPs include six cysteines predicted to form three intraprotease domain disulfide bonds. Enteropeptidase and hepsin each have one and corin has two additional predicted disulfide linkages within the catalytic domain. The presence of an aspartate six residues before the catalytic serine, which in the activated TTSP would be positioned at the bottom of the S1 substrate binding pocket, is indicative that all of the TTSPs have preference for substrates containing an arginine or lysine in the P1 amino acid position (S1 and P1 designations are described (25)). The cleavage specificities and candidate physiological substrates for some of the TTSPs have been elucidated. The predicted cleavage specificity following basic amino acids indicates that the TTSPs are likely to have a degree of autocatalytic activity. Indeed truncated mouse hepsin lacking cytoplasmic and transmembrane domains (16) and the human MT-SP1 proteolytic domain (7) are capable of autoactivation. In contrast, bovine enteropeptidase has extremely low autocatalytic activity (26). Interestingly, the proteolytic domain of bovine enteropeptidase has an additional role in the targeting of enteropeptidase to the apical membrane of enterocytes (27).

Transmembrane Domains—Each of the TTSPs contains a hydrophobic domain near the N terminus. This domain is predicted to span the plasma membrane in such a way that the proteolytic domain lies extracellularly, presumably to localize TTSP proteolytic activity in close proximity to target substrates and/or to permit regulated release of the protein from the cell surface. Cell surface localization has been experimentally demonstrated for enteropeptidase, hepsin (28, 29), MT-SP1 (30, 31), TMPRSS2 (20), and TMPRSS3 (10).

Cytoplasmic Domains—The cytoplasmic domains of the TTSPs (Fig. 1) range in length from 12 amino acids for HAT to 112 amino acids for murine corin. Whether these domains have the potential to support interactions with cytoskeletal components and signaling molecules is not yet known. However, a number of the TTSPs including corin, MT-SP1, st-sb, and TMPRSS2 contain consensus

This minireview will be reprinted in the 2001 Minireview Compendium, which will be available in December, 2001. This work was supported by the National Health and Medical Research Council of Australia, the Queensland Cancer Fund, and the National Institutes of Health.

|| To whom correspondence should be addressed. Tel.: 617-3362-0312; Fax: 617-3362-0107; E-mail: toniA@qimr.edu.au.

¹ The abbreviations used are: ADAM, a disintegrin-like and metalloproteinase; ANP, atrial natriuretic peptide; CUB, C1s/C1r, urokinase growth factor and bone morphogenetic protein 1; ECM, extracellular matrix; HAT, human airway trypsin-like protease; LDL, low density lipoprotein; MAM, matriptase, and receptor protein phosphatase μ ; MT-MMP, membrane-type matrix metalloproteinase; PAI-1, plasminogen activator inhibitor-1; PAR, protease-activated receptor; SEA, sea urchin sperm protein-germkinase-1; SR, Group A scavenger receptor; st-sb, stubble-stubloid; TAPVR, total anomalous pulmonary venous return; TTSP, type II transmembrane serine protease; uPA, urokinase-type plasminogen activator; uPAR, uPA receptor.

² Originally designated TMPRSS3 (10). The Human Genome Nomenclature Committee-approved symbol TMPRSS3 has been allocated to a predicted TTSP-encoding gene located on chromosome 21q22.3 (66). The amino acid sequence of the TMPRSS3 protein has not been reported.

³ Information on the classification and nomenclature of the S1 family of peptidases can be found in the Internet-accessible MEROPS data base.

TABLE I
Summary of type II transmembrane serine proteases

The abbreviations used are: b, brain; bl, bladder; bp, *Drosophila* 36-h pupae; c, colon; de, *Drosophila* 12–18-h embryo; dp, *Drosophila* early prepupae; e, esophagus; h, heart; int, intestine; k, kidney; l, lung; le, leukocytes; li, liver; p, pancreas; pl, placenta; pr, prostate; psi, proximal small intestine (si); s, spleen; st, stomach; t, testes; th, thymus; tr, trachea.

Name	Organism	Other Name	% Identity to Human Orthologue	Accession Number	MW (kDa)	Gene Location	Expression Pattern	mRNA Size (kb)	Reference
Corin	Human	-	100	AF133845	~150 ^{a,c}	4p12-13	h	-	5 (6, 56)
	Mouse	LRP4	82	AB013874	123	5 ^a	h	l, k, l	5 (18)
Enteropeptidase	Human	Enterokinase	100	U09860	158 ^b	21q21	psi	-	4.4 (3)
	Bovine	-	83	U09859 ^a	150 ^b	-	-	-	(14)
	Mouse	-	75	U73378	118.7	-	-	-	(12)
	Rat	-	73	1589367	117.7	-	psi	-	4.4 (13)
	Porcine	-	85	D30799	200 ^a	-	-	-	(15)
MT-SP1	Human	Matrilysin	100	AF133088/AF118224	67 ^b	11q25	c, si, st, pr, l, pl, s, th	k, k, le	3.3 (7, 8, 31)
	Mouse	Epithelin	81	AF042822	94.4	9 ^a	int, k	l, s, th	3 (19)
HAT	Human	-	100	AB002134	48 ^a	-	tr	-	0.9, 1.9, 3.0 (5)
Hepsin	Human	-	100	M18930	51 ^a	19q13.1	li	-	1.8 (4)
	Mouse	-	88	AF030065 ^a	44.7	-	k, li	-	1.8, 1.9 (16)
	Rat	-	88	X70900	44.9	-	-	-	(17)
Stubble-Stubloid	<i>Drosophila</i>	-	-	L11451	85	-	bp, de, dp	-	3.8 (11)
TMPRSS2	Human	-	100	U75329	53.8	21q22.3	pr	-	3.8 (9)
	Mouse	Epithelisin	77	AF113596	53.5	16C2	k	li	1.5, 2.8 (20)
TMPRSS4	Human	-	100	AF179224	68 ^{a,c}	11q23.3	-	-	2.3 (10)

^a Splice variants have been identified. ^b Experimentally derived molecular weight. ^c V5/His₆-tagged protein. ^d Putative assignment based on our unpublished observation that LRP4 sequences have greater than 96% identity with mouse chromosome 5 BAC RP23-294A15 sequences deposited in the GenBank™ htgs database (GenBank™ accession no. AC036146). ^e Closest linkage to the *Flii* gene.

phosphorylation sites for either or both of protein kinase C and casein kinase II. In addition, based on the cellular sorting of other integral membrane proteins (32) it is likely that the cytoplasmic and transmembrane domains also contribute to the targeting of the TTSPs to a particular cell surface in polarized cells.

Stem Regions—The stem regions of the TTSPs contain as many as 11 structural domains that may serve as regulatory and/or binding domains (Fig. 1). These include low density lipoprotein (LDL) receptor class A domains, Group A scavenger receptor (SR) domains, frizzled domains, C1s/C1r, urchin embryonic growth factor and bone morphogenic protein 1 (CUB) domains, sea urchin sperm protein, enterokinase, agrin (SEA) domains, a meprin, A5 antigen, and receptor protein phosphatase μ (MAM) domain, and a disulfide knotted domain. Hepsin is the only TTSP that does not possess an identified structural domain within its stem region. Although functional roles for individual stem region domains have not been demonstrated, the stem region of bovine enteropeptidase has been shown to be required for efficient cleavage of its physiological substrate trypsinogen (26). In addition, the N terminus of the stem region of this protein is required for delivery of enteropeptidase to the apical surface of polarized Madin-Darby canine kidney cells (27).

The most common stem region structural domain is the LDL receptor class A domain: corin contains eight, MT-SP1 four, enteropeptidase two, and TMPRSS2 and TMPRSS4 one each (Fig. 1). Although the function of these domains in the TTSPs has not been demonstrated, in other proteins they bind Ca²⁺ ions and mediate the internalization of macromolecules including serine protease-inhibitor complexes and lipoproteins (33–35). In addition, although LDL receptor domains also function in the uptake of LDLs, increased LDL uptake could not be demonstrated following expression of murine corin in COS cells (18).

Six other structural domains that are thought to be involved in protein-protein interactions or protein-ligand interactions are found in various TTSPs. SR domains (36) are present in corin, enteropeptidase, TMPRSS2, and TMPRSS3; frizzled domains (37) are present in corin; CUB domains (38) are present in enteropeptidase and MT-SP1; SEA domains (39) are present in HAT and enteropeptidase; a MAM domain (40) is present in enteropeptidase; and a disulfide knotted domain (41) is present in st-ab (Fig. 1). In addition to these structural domains, human and mouse MT-SP1s possess a conserved RGD motif (42) present in the first CUB domain. Interestingly, truncated human MT-SP1 lacking cytoplasmic and transmembrane domains remains bound to the cell surface of COS cells (31). Binding may be mediated via an interaction between the MT-SP1 RGD motif and an integrin protein or another

cell surface protein. Alternatively, the mode of attachment could be via a direct link such as a hydrocarbon chain.

Tissue Expression of TTSPs

Although a few of the TTSPs are expressed across several tissue and cell types, in general these enzymes demonstrate relatively restricted expression patterns, indicating that they may have tissue-specific functions (Table I). Enteropeptidase shows a very narrow expression pattern, being restricted in normal tissues to enterocytes of the proximal small intestine (12). Corin expression is also quite specific, with corin mRNA highly expressed in human heart (6) and corin protein expression localized to cardiac myocytes (43). HAT is predominantly expressed in trachea (5, 23). Human TMPRSS2 expression is predominantly associated with prostate (9, 44).⁴ Hepsin, originally identified from liver, is highly expressed in fetal liver and kidney (45). Hepsin mRNA has been reported to be overexpressed by ovarian tumors (46), and protein expression has been localized to tumor cell membranes in renal cell carcinoma (29). TMPRSS4 has only recently been characterized and was identified as a consequence of its strong up-regulation in pancreatic tumors (10). While TMPRSS4 was not detected in normal pancreas, very low level TMPRSS4 mRNA expression was detected in tissues of the gastrointestinal tract and in some tissues of the urogenital tract (10). MT-SP1 was originally identified from a human breast cancer line (30) but shows the broadest pattern of expression of the TTSPs being detected in a wide range of both human (7) and murine tissues (19).

Biochemical Data and Pathophysiological Roles

The majority of the TTSPs have been identified relatively recently and consequently have not been extensively characterized. Enteropeptidase is somewhat of an exception. Although the enzymatic activity ascribed to enteropeptidase was first identified almost a century ago (47) it has been only recently that the complete amino acid sequence was described (3). Enteropeptidase functions near the apex of the digestive enzymatic cascade activating the digestive protease trypsinogen to trypsin, which subsequently activates other enzymes including chymotrypsinogen, proelastase, prolipase, and procarboxypeptidases. Enteropeptidase possesses extremely low autocatalytic activity, and it has been proposed that the serine protease duodenase, secreted by duodenal epitheliocytes, may be its physiological activator (48). Active enteropeptidase con-

⁴ The Northern blot data reported (9) are incorrectly labeled due to inversion of the membranes (Stylianos Antonarakis, personal communication).

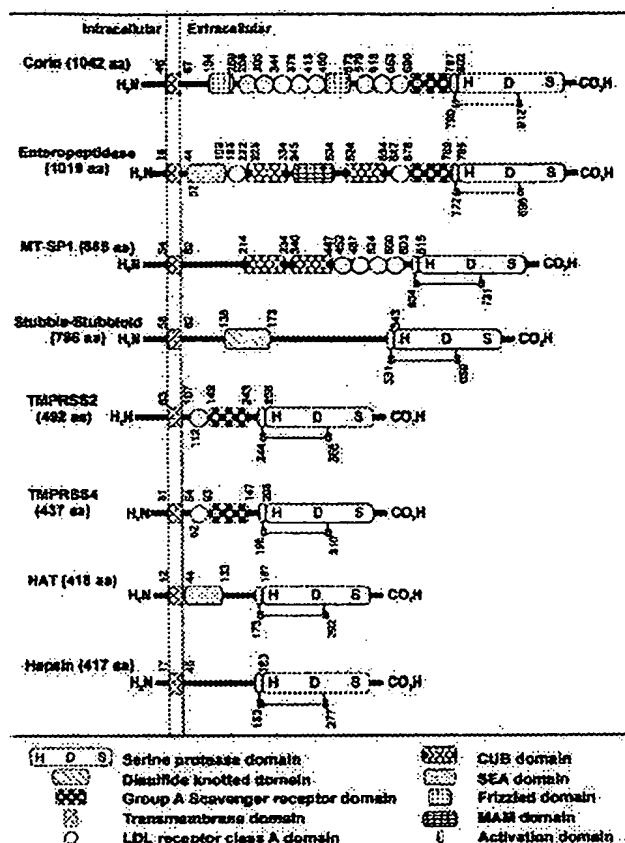


FIG. 1. Type II transmembrane serine protease domain structure. Structures, listed by length, are of the seven human TTSPs and the *Drosophila* TTSP st-sb. The amino acid (aa) sequence of each protein was scanned using the ProfileScan algorithm to confirm the presence of each domain. Numbers delineate the location of each domain.

sists of heavy and light chains that are extensively glycosylated (27, 49). It has recently been reported that physiological concentrations of pancreatic trypsin activate protease-activated receptor (PAR) 2 at the apical membrane of enterocytes (50). PAR2 is a member of the PAR family of signal-transducing, G protein-coupled, plasma membrane-spanning receptors, which are activated by the proteolytic action of select serine proteases (51, 52). These data and the observation that an exosite in the heavy chain of enteropeptidase is required for efficient recognition of trypsinogen (26) suggest that enteropeptidase may play a role in facilitating trypsin-mediated PAR2 activation on enterocytes. Thus enteropeptidase may localize trypsinogen/trypsin at the membrane of enterocytes, initiating a limited proteolytic cascade at the cell surface in close proximity to the trypsin cleavage target PAR2, thereby facilitating receptor activation and signal transduction.

Hepsin is a glycoprotein originally cloned from human liver and hepatoma cell lines and, more recently, implicated in mammalian cell growth and morphology (53), tumor progression (28), and developmental processes, such as blastocyst hatching (16). The importance of hepsin *in vivo*, however, remains unclear as homozygous hepsin null mice are phenotypically normal (54). An as yet unexplained phenotype of the hepsin $-/-$ mice is a 2-fold higher serum concentration of bone-derived alkaline phosphatase compared with wild type mice (55).

The human airway TTSP, HAT, was originally purified as a soluble protein from the sputum of patients with chronic airway diseases. Full-length HAT is synthesized, translocated to the cell surface where it is processed to a soluble form, and then released

from tracheal serous glands as part of the host immune defense system (5).

Significantly, the human heart TTSP, corin, is an *in vitro* activator of pro-atrial natriuretic peptide (ANP), a cardiac hormone essential for the regulation of blood pressure (56), suggesting that corin is the long sought pro-ANP convertase. This proteolytic cleavage is critical for the regulation of ANP activity (57); thus, corin may well prove to be an important factor in the regulation of major cardiovascular diseases. Dysfunctional corin was proposed to be a candidate for the rare congenital heart disease, total anomalous pulmonary venous return (TAPVR), as the corin gene colocalizes to the TAPVR locus on human chromosome 4p12–13 (6). In addition to heart, murine corin is expressed by chondrocytes in a differentiation stage-specific manner during mouse development, suggesting that this protease may play a role during chondrocyte differentiation/bone formation (6). However, while human and murine corin share high homology, common structural features, expression profiles, and syntenic chromosomal locations, these proteases are variant in the lengths of their cytoplasmic domains (45 residues in human and 112 in mouse) and show no conservation in amino acid sequence in this domain. This may indicate that murine and human corin have different but perhaps overlapping species-specific roles, or alternatively the cytoplasmic domain is not essential for corin functions.

In other significant recent experiments it has been shown that MT-SP1 may be involved in initiating signaling and proteolytic cascades via the activation of the cell surface-associated proteins PAR2 and pro-uPA (31). Interestingly, MT-SP1 from breast cancer cells is detected largely as an uncomplexed protein, whereas in milk it is present mainly as a complex with the Kunitz-type serine protease inhibitor hepatocyte growth factor inhibitor-1 (24). It will be important to identify the inhibitor binding domains of MT-SP1 and the function of the protease-inhibitor complex.

TMPRSS2 and TMPRSS4 have been identified through association with cancer. TMPRSS2 is thought to play a role in epithelial cell biology, and its association with prostate carcinogenesis has led to the proposal that it may be a diagnostic or therapeutic target for prostate cancer (44). TMPRSS2 has been proposed to be part of an enzymatic cascade involving the serine proteases prostate-specific antigen and human kallikrein K2 in a manner analogous to the fibrinolytic and blood coagulation cascades (44). TMPRSS4 is over-expressed in pancreatic cancers; however, its functional significance remains unclear (10).

The *Drosophila* serine protease st-sb is one of a number of proteases involved in fly morphogenesis (11) and has a proteolytic function in detaching imaginal disks from extracellular matrices. In addition, the phenotype of st-sb mutants has led to speculation that the encoded protein is involved in outside to inside signal transduction via its cytoplasmic domain, thus resulting in cytoskeletal reorganization and changes in cell shape during morphogenesis (11).

Analogous Membrane-associated Proteolytic Systems

In contrast to the traditional protein catabolic functions of many of the secreted members of the serine protease family and based on the presence of multiple structural domains in the TTSPs, it is tantalizing to speculate that the TTSPs function as key regulators of signaling events at the plasma membrane. Precedents for such functions come from other more well characterized membrane-associated proteolytic systems such as the ADAMs (1), the MT-MMPs (2), and the uPA-uPA receptor system (58).

The ADAMs have recognized and proposed roles in the proteolysis of extracellular matrix (ECM) components and cell surface proteins, in mediating cell adhesion via integrin binding, in cell fusion and signaling via interactions of their cytoplasmic domains, and in RGD-mediated interactions with integrins (59–61). The TTSPs are similarly positioned at the plasma membrane to release ECM components and to proteolytically activate cell surface proteins such as PARs, growth factors, and cytokines, and to interact with cell surface and soluble ligands. In addition, the presence of the cytoplasmic domains indicates that the TTSPs may be capable of interacting with the cytoskeleton and/or with cellular signaling molecules.

The MT-MMPs function in pericellular cascades to activate other MMPs involved in the cleavage of ECM components. The TTSPs may well perform similar functions in activating proteolytic cascades on

the plasma membrane. Indeed, this function has been demonstrated for enteropeptidase in the activation of digestive proteases. Moreover, there is increasing evidence for cross-talk between proteolytic systems. The uPA-uPAR receptor system of cell surface-localized proteolytic activity has a recognized role in the initial stage of MMP activation (62), and other serine proteases are also capable of *in vitro* MMP activation (63, 64). The TTSPs could play a direct role in MMP activation or an indirect role in localizing and activating other serine proteases more directly associated with MMP activation. The activation of uPA by MT-SP1 (31) and subsequent downstream MMP activation could be an example of such cross-talk.

Several other parallels may also be drawn from the uPA-uPAR receptor system. That the TTSPs are directly anchored to the plasma membrane implies that they have potential to mimic localization of the uPA-uPAR system to the leading edge of migrating tumor cells (65). Further, the interaction of the uPA-uPAR system, via a nonproteolytic mechanism, in mediating cell-cell contacts through association with integrins may also parallel TTSP properties. Indeed the multidomain structure of the TTSPs indicates their capacity to interact with multiple partners and suggests the possibility that these membrane proteins may form part of a signalosome-like complex, thereby mediating at the cell surface multiple signaling pathways as is the case for the uPA-uPAR system (58).

Concluding Remarks

What is known about the TTSPs is that they function or have the structural motifs necessary to function as serine proteases. What can be speculated upon is that their numerous and varied nonproteolytic domains are likely to mediate interactions with proteolytic substrates and inhibitors as well as other proteins and ligands. Such interactions will potentially regulate the proteolytic activity of the catalytic domain but perhaps may also have functions quite independent of this domain. Furthermore, given the integral plasma membrane nature of the TTSPs, it is tempting to speculate that at least some of the TTSPs will function directly in transducing signals across the plasma membrane, as has been suggested for the *Drosophila* TTSP st-sb (11). There is clearly a need for a greater understanding of the biology and physiological functions of this group of unique proteases to obtain a better picture of the dynamics occurring on the cell surface. Because of the mosaic structure of the TTSPs it will be important to understand the role of their individual domains as well as the role of each protein *in toto*.

Note Added in Proof—Two cDNAs encoding the putative TTSPs Kesp-2 and XMT-SP1 have recently been identified from *Xenopus laevis* (67).

REFERENCES

- Stone, A. L., Kroeger, M., and Sang, Q. X. (1999) *J. Protein Chem.* **18**, 447–465.
- Seiki, M. (1999) *APMIS* **107**, 137–149.
- Kitamoto, Y., Velle, R. A., Donis-Keller, H., and Sadler, J. E. (1995) *Biochemistry* **34**, 4562–4568.
- Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* **27**, 1067–1074.
- Yamaoka, K., Masuda, K., Ogawa, H., Takagi, K., Umamoto, N., and Yasuoka, S. (1998) *J. Biol. Chem.* **273**, 11895–11901.
- Yan, W., Sheng, N., Seto, M., Morser, J., and Wu, Q. (1999) *J. Biol. Chem.* **274**, 14926–14935.
- Takeuchi, T., Shuman, M. A., and Craik, C. S. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11054–11061.
- Lin, C.-Y., Anders, J., Johnson, M., Sang, Q. A., and Dickson, R. B. (1999) *J. Biol. Chem.* **274**, 18231–18236.
- Paoloni-Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C., and Antonarakis, S. E. (1997) *Genomics* **44**, 309–320.
- Wallrapp, C., Hahnel, S., Muller-Pillasch, F., Burghardt, B., Iwamura, T., Ruthenburger, M., Lerch, M. M., Adler, G., and Gress, T. M. (2000) *Cancer Res.* **60**, 2602–2606.
- Appel, L. F., Prout, M., Abu-Shumays, R., Hammonds, A., Garbe, J. C., Fristrom, D., and Fristrom, J. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90**, 4937–4941.
- Yuan, X., Zheng, X., Lu, D., Rubin, D. C., Pung, C. Y., and Sadler, J. E. (1998) *Am. J. Physiol.* **274**, G342–G349.
- Yahagi, N., Ichinose, M., Matsushima, M., Matsubara, Y., Miki, K., Kurokawa, K., Fukumachi, H., Tashiro, K., Shiokawa, K., Kageyama, T., Takahashi, T., Inoue, H., and Takahashi, K. (1996) *Biochem. Biophys. Res. Commun.* **219**, 806–812.
- Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W., and Sadler, J. E. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91**, 7588–7592.
- Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., Umayama, H., Inoue, H., Takahashi, T., and Takahashi, H. (1994) *J. Biol. Chem.* **269**, 19976–19982.
- Vu, T. K. H., Liu, R. W., Haaksma, C. J., Tomasek, J. J., and Howard, E. W. (1997) *J. Biol. Chem.* **272**, 31315–31320.
- Farley, D., Raymond, F., and Nick, H. (1993) *Biochim. Biophys. Acta* **1173**, 350–352.
- Tomita, Y., Kim, D. H., Magori, K., Fujino, T., and Yamamoto, T. T. (1998) *J. Biochem. (Tokyo)* **124**, 784–789.
- Kim, M. G., Chen, C., Lyu, M. S., Cho, E. G., Park, D., Kozak, C., and Schwartz, R. H. (1999) *Immunogenetics* **49**, 420–428.
- Jacquinet, E., Rao, N. V., Rao, G. V., and Hoidal, J. R. (2000) *FEBS Lett.* **468**, 93–100.
- Louvard, D., Maroux, S., Baratti, J., and Desnuelle, P. (1973) *Biochim. Biophys. Acta* **309**, 127–137.
- Fonseca, P., and Light, A. (1983) *J. Biol. Chem.* **258**, 14516–14520.
- Yasuoka, S., Ohnishi, T., Kawano, S., Tsuchihashi, S., Ogawara, M., Masuda, K., Yamaoka, K., Takahashi, M., and Sano, T. (1997) *Am. J. Respir. Cell Mol. Biol.* **16**, 300–308.
- Lin, C.-Y., Anders, J., Johnson, M., and Dickson, R. B. (1999) *J. Biol. Chem.* **274**, 18237–18242.
- Schechter, I., and Berger, A. (1967) *Biochem. Biophys. Res. Commun.* **27**, 157–162.
- Lu, D., Yuan, X., Zheng, X., and Sadler, J. E. (1997) *J. Biol. Chem.* **272**, 31293–31300.
- Zheng, X., Lu, D., and Sadler, J. E. (1999) *J. Biol. Chem.* **274**, 1596–1605.
- Kazama, Y., Hamamoto, T., Foster, D. C., and Kisiel, W. (1995) *J. Biol. Chem.* **270**, 66–72.
- Zacharski, L. R., Ornstein, D. L., Memoli, V. A., Rousseau, S. M., and Kisiel, W. (1998) *Thromb. Haemostasis* **78**, 876–877.
- Lin, C.-Y., Wang, J. K., Torri, J., Dou, L., Sang, Q. A., and Dickson, R. B. (1997) *J. Biol. Chem.* **272**, 9147–9152.
- Takeuchi, T., Harris, J., Huang, W., Yan, K. W., Coughlin, S. R., and Craik, C. S. (2000) *J. Biol. Chem.* **275**, 26333–26342.
- Keller, P., and Simons, K. (1997) *J. Cell Sci.* **110**, 3001–3009.
- Brown, M. S., Herz, J., and Goldstein, J. L. (1997) *Nature* **388**, 629–630.
- Nykjer, A., Conese, M., Christensen, E. I., Olson, D., Cremona, O., Glieman, J., and Blasi, F. (1997) *EMBO J.* **16**, 2610–2620.
- Kounnas, M. Z., Church, F. C., Argraves, W. S., and Strickland, D. K. (1996) *J. Biol. Chem.* **271**, 6523–6529.
- Resnick, D., Chatterton, J. E., Schwartz, K., Slayter, H., and Krieger, M. (1996) *J. Biol. Chem.* **271**, 26924–26930.
- Cadigan, K. M., and Nusse, R. (1997) *Genes Dev.* **11**, 3286–3305.
- Bork, P., and Beckmann, G. (1993) *J. Mol. Biol.* **231**, 539–545.
- Bork, P., and Pathay, L. (1995) *Protein Sci.* **4**, 1421–1425.
- Beckmann, G., and Bork, P. (1993) *Trends Biochem. Sci.* **18**, 40–41.
- Muta, T., Hashimoto, R., Miyata, T., Nishimura, H., Toh, Y., and Iwanaga, S. (1990) *J. Biol. Chem.* **265**, 22426–22433.
- Hynes, R. O. (1992) *Cell* **69**, 11–25.
- Hooper, J. D., Scarman, A. L., Clarke, B. E., Normyle, J. F., and Antalis, T. M. (2000) *Eur. J. Biochem.* **267**, 6931–6937.
- Lin, B., Ferguson, C., White, J. T., Wang, S., Vessella, R., True, L. D., Hood, L., and Nelson, P. S. (1999) *Cancer Res.* **59**, 4180–4184.
- Tsuji, A., Torres-Rosado, A., Arai, T., Le Beau, M. M., Lemons, R. S., Chou, S. H., and Kurachi, K. (1991) *J. Biol. Chem.* **266**, 16948–16953.
- Tanimoto, H., Yan, Y., Clarke, J., Korourian, S., Shigemasa, K., Parmley, T. H., Parham, G. P., and O'Brien, T. J. (1997) *Cancer Res.* **57**, 2884–2887.
- Pavlov, I. P. (1902) *The Work of the Digestive Glands*, 1st Ed., pp. 148–163, translated by W. H. Thompson, Charles Griffin & Co., London.
- Zamolodchikova, T. S., Sokolova, E. A., Lu, D., and Sadler, J. E. (2000) *FEBS Lett.* **466**, 295–299.
- Lu, D., and Sadler, J. E. (1998) in *Handbook of Proteolytic Enzymes* (Barrett, A. J., Rawlings, N. D., and Woessner, J. F., eds) pp. 50–54, Academic Press Ltd., London.
- Kong, W., McConlogue, K., Khitin, L. M., Hollenberg, M. D., Payan, D. G., Bohm, S. K., and Bunnett, N. W. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94**, 8884–8889.
- Dery, O., and Bunnett, N. W. (1999) *Biochem. Soc. Trans.* **27**, 246–254.
- Hollenberg, M. D. (1999) *Trends Pharmacol. Sci.* **20**, 271–273.
- Torres-Rosado, A., O'Shea, K. S., Tsui, A., Chou, S. H., and Kurachi, K. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7181–7185.
- Wu, Q., Yu, D., Post, J., Halks, M. M., Sadler, J. E., and Morser, J. (1998) *J. Clin. Invest.* **101**, 321–326.
- Kawamura, S., Kurachi, S., Deyashiki, Y., and Kurachi, K. (1999) *Eur. J. Biochem.* **262**, 755–764.
- Yan, W., Wu, F., Morser, J., and Wu, Q. (2000) *Proc. Natl. Acad. Sci. U. S. A.* **97**, 8525–8529.
- Lang, R. E., Tholken, H., Ganten, D., Luft, F. C., Ruskoaho, H., and Unger, T. (1985) *Nature* **314**, 264–266.
- Kashelnick, Y., Ehart, M., Stockinger, H., and Binder, B. R. (1999) *Thromb. Haemostasis* **82**, 305–311.
- Wolfsberg, T. G., Primakoff, P., Myles, D. G., and White, J. M. (1995) *J. Cell Biol.* **131**, 275–278.
- Wolfsberg, T. G., and White, J. M. (1996) *Dev. Biol.* **180**, 389–401.
- Schlondorff, J., and Blobel, C. P. (1999) *J. Cell Sci.* **112**, 3603–3617.
- Carmeliet, P., Moons, L., Lijnen, R., Baes, M., Lemaire, V., Tipping, P., Drew, A., Eeckhout, Y., Shapiro, S., Lupu, F., and Collen, D. (1997) *Nat. Genet.* **17**, 439–444.
- Nagase, H., Englund, J. J., Suzuki, K., and Salvesen, G. (1990) *Biochemistry* **29**, 5783–5789.
- Ramos-DeSimone, N., Hahn-Dantona, E., Siple, J., Nagase, H., French, D. L., and Quigley, J. P. (1999) *J. Biol. Chem.* **274**, 13066–13076.
- Blasi, F. (1999) *Thromb. Haemostasis* **82**, 298–304.
- Hattori, M. et al. (2000) *Nature* **405**, 311–319.
- Yamada, K., Takabatake, T., and Takeshima, K. (2000) *Gene (Amst.)* **252**, 209–216.

Exhibit 16

Cloning, genomic organization, chromosomal assignment and expression of a novel mosaic serine proteinase: epitheliasin

Eric Jacquinet, Narayanam V. Rao, Gopna V. Rao, John R. Hoidal*

Department of Internal Medicine, Division of Respiratory, Critical Care and Occupational Medicine, Pulmonary Division, Winthrope Building, Rm. 743A, 50N. Medical Drive, University of Utah Health Sciences Center and VA Medical Center, Salt Lake City, UT 84132, USA

Received 20 January 2000

Edited by Horst Feldmann

Abstract We report the isolation of a cDNA encoding a novel murine serine proteinase, epitheliasin. The cDNA spans 1753 bp and encodes a mosaic protein with a calculated molecular mass of 53 529 Da. Its domains include a cytoplasmic tail, a type II transmembrane domain, a low-density lipoprotein receptor class A domain, a cysteine rich scavenger receptor-like domain and a serine proteinase domain. The proteinase portion domain shows 46–53% identity with mouse neurotrypsin, acrosin, hepsin and enteropeptidase. The gene, located in the telomeric region in the long arm of mouse chromosome 16, consists of 14 exons and 13 introns and spans approximately 18 kb. Epitheliasin is expressed primarily in the apical surfaces of renal tubular and airway epithelial cells.

© 2000 Federation of European Biochemical Societies.

Key words: Serine proteinase; Mosaic protein; Epitheliasin

1. Introduction

Proteinases are implicated in a wide spectrum of physiologic and pathophysiological processes in the kidney. Renin, a proteinase synthesized in renal cortical cells plays a major role in the regulation of blood pressure and electrolyte balance by converting angiotensinogen to angiotensin I. Furthermore, the renal kallikrein-kinin system activated under conditions of mineralocorticoid excess represents a compensatory response against the development of hypertension and renal injury induced by salt excess. Proteolytic enzymes also have been ascribed important roles in both leukocyte-dependent and independent models of glomerular diseases (reviewed in [1]). Recently, Vallet and colleagues identified a novel serine proteinase from *Xenopus laevis* kidney epithelial cells. CAP 1, involved in activation of the epithelial sodium channel. EnaC [2]. This was the first report of channel activating activity of an endogenous proteinase.

In the present report, we describe a novel serine proteinase expressed in murine renal epithelial cells with sequence homology to CAP1. The enzyme, that we term epitheliasin, is a modular protein consisting of five sequence motifs, a cytoplasmic tail, a type II transmembrane (TM) domain, a low-density lipoprotein receptor class A (LDLRA)-like domain, a cysteine rich scavenger receptor-like (SRCR) domain and a serine proteinase domain. The sequence and structural features of epitheliasin cDNA and gene, its chromosomal localization and

tissue expression are described. Epitheliasin has sequence identity to a human cDNA recently cloned by exon trapping named TMPRSS2 [3]. However, the tissue distribution of epitheliasin and TMPRSS2 is strikingly different.

2. Materials and methods

2.1. Materials

Multiple tissue Northern blots, ExpressHyb hybridization solution, rapid amplification of cDNA ends (RACE) ready cDNAs from mouse kidneys and Marathon cDNA kits were from CLONTECH (Palo Alto, CA, USA). TA cloning kits were from Invitrogen (Carlsbad, CA, USA). LA PCR kits were from Panvera (Madison, WI, USA). Klenow DNA polymerase, [α - 32 P]dCTP (3000 Ci/mmol) and [γ - 32 P]dATP (3000 Ci/mmol) were from Amersham Life Science (Arlington Heights, IL, USA). BUPHTM Tris-glycine SDS, Tris-glycine and Immunogen Conjugation kits were from Pierce (Rockford, IL, USA). Alkaline phosphatase conjugated goat anti-rabbit antibody was from Zymed (San Francisco, CA, USA). BCIP/NBT tablets were from Sigma (St Louis, MO, USA). Citra solution and VIP substrate were from Vector Laboratories (Burlingame, CA, USA). Blocking reagent, SA-HRP and biotinyl tyramide were supplied by NEN Life Science Products (Boston, MA, USA).

2.2. Identification and cloning of epitheliasin cDNA

A conserved sequence around the serine active site residue (GGIDSCQGDSSGGPLVC) was used to search the mouse EST database using TBLASTn. Of the 100 ESTs initially identified, a novel EST (ub58g01.s1) containing 389 nt and its mirror sequence (ub58g01.r1) were further analyzed using the non-redundant databases, BLASTn and BLASTx. Four overlapping sequences were found from these searches, one was from a kidney library (uc81c11.y1), two from a mammary gland library (vf86g09.r1, ve37e12.r1), and one from a blastocyst library (v164c03.r1).

To obtain the full-length cDNA of interest the RACE strategy was employed. Initially, LA PCR was utilized to amplify mouse kidney cDNA employing a sense primer (5'- 36 CCATACTGAACCTCTC-ATGCTGCT- 13 -3') designed based on the novel sequence and an anchor primer, AP1. The initial PCR product was subjected to nested PCR using a sense (5'- 14 CTGACACAGGCAGGATGGCATTG 9 -3') and an anti-sense primer (5'- 1455 GTGGATTAGCTGTTCGCC-CTCATT 1478 -3'). This nested reaction amplified a 1.5 kb product that was ligated into the pCR $^{3.1}$ vector and sequenced using an ABI automatic sequencer.

To obtain the 3' end, mouse kidney cDNA was subjected to 3'-RACE. The cDNA was amplified using AP1 and a sense primer (5'- 36 CCATACTGAACCTCTCATGCTGCT- 13 -3'). The product was diluted (1:50) and a nested PCR amplification was performed using a second anchor primer, AP2, and a sense primer (5'- 14 CTGACACA-GGCAGGATGGCATTG 9 -3'). The 2 kb PCR product obtained was cloned and sequenced as described above.

2.3. Genomic cloning and analysis

To obtain the epitheliasin gene, a mouse genomic bacterial artificial chromosome (BAC) library (Genome Systems, St Louis, MO, USA) was screened using a 0.7 kb probe extending from 831 to 1477 nt of mouse epitheliasin cDNA. A single clone (BAC-24) was identified and confirmed by sequencing to contain the entire epitheliasin gene. To

*Corresponding author. Fax: (1)-801-585 3355.
E-mail: jhoidal@med.utah.edu

identify the intron junction borders. DNA from BAC-24 was directly sequenced using oligonucleotide primers defined initially by the cDNA sequences and subsequently by derived sequences. Southern analysis was used to determine the size of the epitheliasin gene.

2.4. Chromosomal assignment

The plasmid clone (BAC-24) obtained from the genomic library was used as a probe for chromosomal localization by fluorescence in situ hybridization (FISH). The probe was nick translation-labeled with biotin, hybridized to metaphase chromosomes and detected with Cy-3-conjugated streptavidin. Chromosome spreads were prepared by standard procedures and G-banded after trypsin treatment and Wright's staining. Hybridization and detection conditions on metaphase chromosomes were performed as previously described [4]. Probe signals were detected with the Cy3 conjugate viewed using an epifluorescence microscope. The fluorescence image was overlaid on the G-banded image to localize the gene.

2.5. Northern blot analysis

Mouse multi-tissue blots containing 2 µg of poly(A) RNA in each lane were prehybridized for 1 h at 68°C, then hybridized at 68°C with a 1.5 kb [α - 32 P]dCTP-labeled probe that represented the coding region of the mouse epitheliasin cDNA. After low stringency washes, the blots were washed at high stringency at 50°C and autoradiographed.

2.6. Production of antibodies against epitheliasin

Rabbit polyclonal antiserum was raised to a synthetic peptide, cS¹¹²HPNYDSKTKNND²⁴³, located in the serine proteinase region of epitheliasin. The peptide was chosen based on predicted surface hydrophilicity and antigenicity. The peptide was coupled to keyhole-limpet hemocyanin. Subcutaneous injections were given to rabbits with 100 µg of conjugate that was emulsified in Freund's complete adjuvant and then boosted with the same amount of antigen in Freund's incomplete adjuvant at 2 week intervals until a titer of >1:4000 was obtained. The presence of anti-peptide antibodies was assessed by dot blot analysis using the peptide linked to ovalbumin as the antigen.

2.7. Immunohistology

Mouse kidneys and lungs were fixed in buffered 10% formaldehyde, and embedded in paraffin. Sections were cut at 5 µm depths, deparaffinized and rehydrated. Following antigen retrieval performed with 1× Citra solution in a microwave oven for 15 min at 700–900 W, the samples were washed in PBS. Endogenous peroxidase activity was blocked with 20% methanol and 3% H₂O₂ in PBS for 30 min at room temperature. The tissue was permeated using 10% Triton X-100 in PBS for 20 min at room temperature. Endogenous biotin was blocked by Vector Block avidin solution for 30 min at room temperature followed by Vector Blocking solution for 30 min at room temperature. The sections were then incubated with epitheliasin peptide anti-serum, dilution 1/500 in Block solution overnight at 4°C in a humid chamber. After washing with TNT, 1/500 horse anti-rabbit IgG serum in TNT was applied for 30 min at room temperature. The slides were then incubated with 1/100 SA-HRP in TNT for 30 min at room temperature. The signal was amplified with biotinyl tyramide for 5 min at room temperature. This was followed by a re-incubation with 1/100 SA-HRP in TNT. The signal was visualized using VIP substrate solution. The same process was applied to the slides used as controls, but epitheliasin anti-serum was replaced by non-immune rabbit serum.

3. Results and discussion

3.1. Cloning and analysis of the epitheliasin full-length cDNA

Fig. 1 shows the nucleic acid and deduced amino acid sequences of the complete cDNA reconstituted from the RACE fragments. As demonstrated by the immunohistochemistry described in a following section, the encoded protein is highly expressed in epithelial tissue. Accordingly, we named the protein epitheliasin. The composite cDNA spans 1753 nt. A 5' untranslated region (UTR) extends 100 nt. The first in-frame ATG (1–3 nt) was assigned as the codon for the Met trans-

lation initiator since the sequence around this codon (ACGATGG) conforms to the Kozak consensus sequence for mammalian protein biosynthesis [5]. A single open reading frame begins with the ATG and extends 1470 nt. This is followed by a stop codon, TAA (1471–1473 nt) and a 3' UTR of 152 nt, terminating in a poly(A)+tail of 28 nt. A consensus polyadenylation site (ATTAAA, 1600–1605 nt) is located 20 nt upstream of the poly (A)+tail.

3.2. Characteristics of the sequence and structural features of epitheliasin

The open reading frame encodes a protein of 490 amino acids with a calculated molecular mass of 53 529 kDa. Comparisons with sequences in GenBank, EMBL and SWISS PROT reveal that the epitheliasin cDNA encodes a multidomain serine proteinase. A typical amino-terminal signal sequence is not present, but a hydrophobic region is present near the amino terminus (Leu⁸⁴ to Trp¹⁰⁵). This 22 amino acid region is flanked by charged amino acids (Lys and Arg) and corresponds to a transmembrane domain [6]. Based on the difference in total charge between the 15-residue sequences on either side of the membrane-spanning domain epitheliasin can be classified as a type II integral membrane bound protein [7,8] that has a cytosol facing amino-terminal tail region consisting of 83 amino acids (Met¹ to Ser⁸³) and an extracellular facing COOH-terminal modular region. The absence of a signal peptide and the presence of a transmembrane domain in epitheliasin are analogous to homologous serine proteinases, enteropeptidase, a key enzyme in digestion that is responsible for the conversion of trypsinogen to trypsin [9], hepsin, a membrane-associated proteinase involved in the formation of thrombin on cell surfaces [10], and a recently described human airway trypsin-like proteinase [11].

The predicted domain structure of epitheliasin is shown in Fig. 2. A LDLRA domain extending from Cys¹¹² to Cys¹⁴⁷ and containing six cysteines follows the transmembrane domain. This domain motif is found in a number of proteins that are functionally unrelated to the LDLR family, including clotting proteinases and enteropeptidase. In each of these proteins the domain is thought to function as a protein-binding domain. The LDLRA domain in epitheliasin is similar to other typical LDLRA domains that are about 40 amino acids long and contain six cysteines [12]. The cysteines form intradomain bridges resulting in a cluster of negatively charged residues in a single loop positioned for high affinity binding to positively charged sequences in LDLR ligands.

Following the LDLRA domain, an SRCR-like domain extends from Val¹⁴⁸ to Gly²⁴³. SRCR domains are classified into two groups, group A and B according to the number of conserved cysteine residues, six or eight, respectively [13]. In a recent analysis, all but one of the 33 independent SRCR domains that had been previously identified had six or eight cysteines [14]. An unusual feature of this domain in epitheliasin is that it contains only four cysteines. These cysteine residues in epitheliasin are completely conserved in position suggesting that the domain belongs to group A. The SRCR domain that is closest to that in epitheliasin is present in complement factor I (CFI), a serum proteinase that regulates the complement cascade by cleaving C3b and C4b. CFI contains a single SRCR domain with five cysteines [13].

The function of SRCR domains is largely unknown. It seems likely that most of these domains are involved in

100 GAAGCAAGA GCTCGACAG AGGCGGACAG GGGCGACAC GGNACAGTC AACTACAGCA AGCCCATAC TGAATCTCT ATCTCTCTA CACGCGAG ATG GCA TTG AAC TCA GCG
119 TCA CCT CCA GGA ATC GGA CCT TGC TAT GAG AAC CAC GGG TAT CAG TCT CAG CAC ATC TCT CCT CCG AGA CCA CCA GTC GCT CCC TAT AAC TTG TAT
S P P G I G P C Y E N H G Y Q S E H I C P P P P V A P N G Y N L Y
121 CCA GCC CAG TAC CCA TCT CCA GTG CCT CAG TAT GCT CCG AGG ATT ACA AGC CAA GGC TCA ACA TCT GTC ATC ACA CAT CCC AAG TCC TCA GGA GCA
P A Q Y Y P S P V P Q Y A P R I T T Q A S T S V I H T H P K S S G A
123 CCG TGC ACC TCA AAG TCT AAG AAA TCG CTG TGT TTA GCC CTT GCC CTG GGC ACT GTC CTC AGC GGA GCT GCT GTC TGT GTC CTT TCG AGG TTC TCG
P C T S K S K K S L C L A L A L G T V L T G A A V A A V L L W R P W
125 GAC AGC AAC TGT TCT ACG TCT GAG ATG GAG TGT GGG TCT CTA GGC ACA TGC ATC AGC TCT TCT CTC TCG TGT GAC GGC GTA GCA CAT TGT CCC AAC GGA GAA
D S N C S T S E M E C G S L G T C I S S L W C D G V A H C P N G B
127 GAT GAG AAC COT TGT GGT CTC TAC GCA CAA AGC TTC ATC CTC CAG GTT TAC TCA TCT CAG AGG AAA GCC TGG TAT CCC GTG TCG CAG GAT GAT TGG AGT
D E N R C V R L Y G Q S F I L Q V Y S S Q R K A W Y P V C Q D D W S
129 GAG AAC TAC GGG AGA GCA GCA TGT AAA GAC ATG GGA TAC AAG AAC AAT TTT TAT TCC AGC CAA GGG ATA CCA CAC CAG AGC GGG GCA AGC AGC TTT ATG AAG
E N Y G R A A C K D M G Y K N N P Y S S Q G I P D Q S G A T S P H R
131 CTG AAT GTG AGC TCA GGC AAT GTT GAC CTC TAT AAA AAA CTC TAC CAC ACT GAC TCA TGT TCA TCC CCG ATG GTG GTT TCT TCG CCG TGT ATA GAA TCC GGG
L N V S S G N V D L Y K K L Y H S D S C S S R H V V S L R C I E C G
133 GTT CCG TCA GTC AAA CCG CAG AGC AGG ATT GTG GGT GGA TTG ANT GCC TCA CCA GGA GAC TGG CCC TGG CAG GTC AGC CTG CAC GTC CAA GGC GTC CAC GTC
V R S V K R Q S R I V G G L N A S P G D W P W Q V S L H V Q G V H V
135 TGC GGA GGC TCC ATC ACC CCC GAG TCG ATT GTG AGC GGC CCC CAC TGT GTG GAA GAA CCC CTC AGC GGC CCG AGG TAC TCG ACG GCA TTT CCG GGA ATT
C G G S I I T P E W I V T A A H C V E E P L S G P R Y W T A P A G I
137 CTG AGA CAG TCT CTC ATG TTC TAT GGA AGT AGA CAC CAG GTA GAA AAA GTA ATT TCC CAT CCA AAT TAC GAC TCT AAG ACC AAG AAT AAC GAC ATT GCT CTC
L R Q S L M F Y G S R H Q V E K V I S H P N Y D S K T K N N D I A L
109 ATG AAG CTG CAG ACA CCT TTG GCT TTT AAT GAT CTA GTG AAG CCA CTA GTC TGT CTG CCG AAC CCA GGC ATG ATG CTA GAC CTA GAC CAG GAA TCC TGG ATT TCG
M K L Q T P L A P N D L V K P V C L P N P C M L D L D Q E C W I S
111 GGG TGG GGG ACC TAT CAG AAA GGG AAG ACC TCG GAC GTG TTT AAT GCT GCC ATG GTA CCC TTG ATC CAG CCC TCC AAA TGT AAT AGT AAA TAC ATA TAC
G W G A T Y E K G K T S D V L N A A H V P L I E P S K C N S K Y I Y
113 AAC AAC CTA ATC ACA GCC ATG ATC TGT GGC GGC TTC CTC CAG GGG TCT GTC GAC TCT TGC CAG GGA GAC AGT GGA GGC CCG CTG GTT ACT TTG AAG AAT
N N L I T P A H I C A G F L Q G S V D S C Q G D S G G P L V T L K N
115 GGG ATC TGG TGG CTG ATT GGG GAC AGC AGC TGG GGC TGT GCC AAG CCA CTC AGA CCT GGA GTA TAC GGG AAC GTG ACG GTA TTT ACA GAT TGG ATC
G I W W L I G D T S W G S G C A K A L R P G V Y G N V T V P T D W I
117 TAC CAG CAA ATG AGC GCG AAC AGC TAA TCCACATGCG TTGTGCGCAG ACTTCTCTTC TCTTCAACAA CCTTTTGCAA GAAACACGAG GGGCTGATTT TTAATCTTCT GTCCACATG
Y Q Q M R A N S
119 TACCTTTTGA GATGATTCGA AGGGCTTTTC ACTTTTATTA AACAGTACT TGTGTGACTG TGAAGAAAAA AAAAAA
121

Fig. 1. Nucleic acid and deduced amino acid sequences of full-length cDNA encoding epitheliasin. Nucleic acid numbering starts at the A nt of the putative translation initiation codon (ATG), with positive and negative numbers preceding to downstream and upstream of sequence, respectively. The consensus translation initiation codon encompassed by ribosomal binding site sequence and putative polyadenylation signal are underlined. An asterisk shows the termination codon. Amino acid residues in single letter code are numbered starting at the first Met residue in the open reading frame and the numbers are shown on the right end of each line. Potential N-linked glycosylation sites are boxed and the encircled amino acid residues are those of the catalytic triad.

ing to molecules on the cell surface or in the extracellular space. Direct evidence supporting the idea that SRCR domains mediate binding to other cell surface proteins or extracellular proteins has recently been provided [14,15].

3.3. Features of serine proteinase domain

The proteinase domain begins with Ile²⁵⁴ and represents the major domain (about 50%) of the encoded protein. The predicted molecular mass of the domain is 25 892 kDa. The domain contains all the major features common to the S1 family of the chymotrypsin (or SA) clan of serine proteinases. The

residues contributing to the salient structural features in chymotrypsin include: (1) His⁵⁷, Asp¹⁰², and Ser¹⁹⁵ that make up the catalytic triad, (2) Gly¹⁹³, Asp¹⁹⁴ and Ser¹⁹⁵ that form an oxyanion hole required for catalytic efficiency, (3) Ser²¹⁴, Trp²¹⁵ and Gly²¹⁶ that bind the main-chain of a substrate and (4) residues that occupy the bottom (Ser¹⁸⁹) and sides (Gly²¹⁶ and Gly²²⁶) of the substrate specificity pocket (S₁ subsite). All of the residues contributing to the first three features and the residues Gly²¹⁶ and Gly²²⁶ on the sides of the substrate specificity pocket of chymotrypsin are strictly conserved in epitheliasin. However, in epitheliasin the residue corre-

Table 1
Exon-intron junctions organization of epitheliasin gene

3' splice site	Exon size in Amino acid	5' splice site	Phase
-CA CAG GTGAGAAGCGCGCCG			
GTTCCTTCCTTCAG GTC ACC ⁴⁴ (5)	AAC TCA N ⁴ S ⁵	0
TTTCCATTGTTTAG GGG TCA G ⁶ S ⁷ (73)	ACC TCA A T ⁷⁷ S ⁷⁸	0
CTTTCTTCCCGCAG AG TCT K ⁷⁹ S ⁸⁰ (29)	AGG TTC T R ¹⁰⁶ P ¹⁰⁷	I
CCAATACAATGCCAG GG GAC W ¹⁰⁸ D ¹⁰⁹ (39)	AAC CG N ¹⁴³ R ¹⁴⁴	I
TTCTTCTCCTTCAG T TGT GTT C ¹⁴⁷ V ¹⁴⁸ (44)	TAC AA Y ¹⁸⁹ K ¹⁹⁰	I
TGTCTTTTTTCCAG G AAC AAT N ¹⁹³ N ¹⁹² (37)	CAC AG H ²²⁶ S ²²⁷	II
CTTTTCTTTTCCAG T GAC TCA D ²²⁸ S ²²⁹ (14)	TGT ATA G C ²⁴⁰ I ²⁴¹	II
GCTTGTCACCTCAG AA TGC E ²⁴² C ²⁴³ (57)	GAA GA E ²⁸⁷ E ²⁸⁸	I
CTTCTGTCTCTCAAG A CCC CTC P ²⁸⁹ L ³⁰⁰ (58)	TTT AAT G F ³³⁵ N ³³⁶	II
CTCTTCTTTAAACAG AT CTA D ³³⁷ L ³³⁸ (32)	GAG AAA G E ³⁸⁷ K ³⁸⁸	I
TGCCTCTGTTGTTAG GG AAG G ³⁸⁹ K ³⁹⁰ (48)	TGC CAG C ⁴³³ Q ⁴³⁴	I
TGCTGTGTTCCCCAG GGA GAC G ⁴³⁷ D ⁴³⁸ (51)	ATG AGG H ⁴⁸⁶ R ⁴⁸⁷	0
TTCTTATTTGCACAG GCG AAC A ⁴⁸⁸ N ⁴⁸⁹ (3)		0

Fig. 2. T
domain.
refer to

spondin
residue.
cleavage
specifici
Comy
teinase
indicate
mouse
and neu
chymotr
substrat
pus laev
epithelia
Based
predict
that is c
the Arg
the enzy
proteoly
or the c
in epithe
sin is sy
intracell
prior to
Arg-Gln
lle-Val-C
teinase
sequence
sin are
suggesti
in proce
Based
with oth
dict that
two cha
domain.

Fig. 3. Sc
of the mo

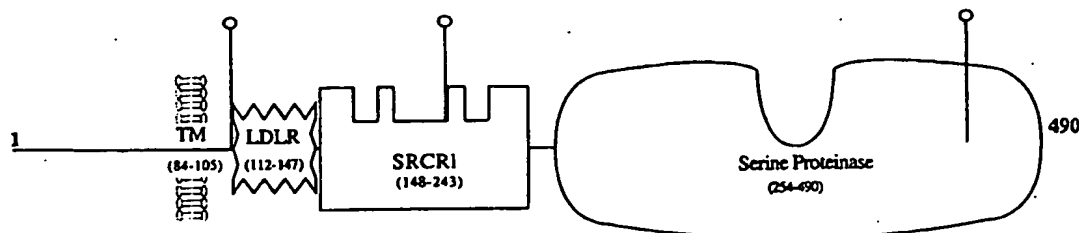


Fig. 2. The domain organization of epitheliasin. Starting at the NH₂-terminus the epitheliasin contains a TM domain followed by a LDLRA domain, a SRCR domain, and finally the serine proteinase domain. N-glycosylation sites are indicated by a circle. The numbers in parentheses refer to the amino acid residues of each domain.

responding to Ser¹⁸⁹ of chymotrypsin is replaced by an acidic residue, Asp. This suggests that epitheliasin has specificity for cleavage after Lys or Arg, indicating a trypsin-like substrate specificity for the enzyme.

Comparison of the amino acid sequence encoding the proteinase domain in epitheliasin with other serine proteinases indicates that this region of the protein shares identity with mouse enteropeptidase (53%), hepsin (51%), acrosin (48%), and neurotrypsin (46%), all multi-domain members of the chymotrypsin family of serine proteinases with trypsin-like substrate specificity. The aforementioned CAP 1 from *Xenopus laevis* kidney epithelial cells has a sequence identity with epitheliasin of 44%.

Based on findings with related vertebrate trypsinogens we predict that epitheliasin is synthesized as an inactive zymogen that is converted to an active serine proteinase by cleavage of the Arg²⁵³-Ile²⁵⁴ peptide bond in the extracellular domain of the enzyme. Most vertebrate trypsinogens are activated by proteolytic cleavage of a Lys (Arg)-Ile bond. The identity or the origin of the proteinase responsible for this cleavage in epitheliasin is not known. One possibility is that epitheliasin is synthesized as a single-chain zymogen and undergoes intracellular cleavage and activation by a furin-like enzyme prior to insertion into the membrane. This is based on the Arg-Gln-Ser-Arg²⁵³ sequence that immediately precedes the Ile-Val-Gly-Gly²⁵⁷ representing the NH₂-terminus of the proteinase domain. Arg-X-X-Arg motifs are furin recognition sequences [16–20]. Interestingly, all the domains of epitheliasin are flanked by recognition sites for furin-like enzymes, suggesting the need to clarify the role of furin-like enzymes in processing of epitheliasin.

Based on the structure of enteropeptidase and a comparison with other chymotrypsin-like serine proteinases, we also predict that epitheliasin, following intracellular cleavage, forms two chains with the smaller chain containing the proteinase domain, and the larger the membrane-spanning segment, and

the LDLRA and SRCR-like domains that may serve as substrate recognition sites. Several chymotrypsin-like serine proteinases including enteropeptidase have a disulfide bond that covalently links the two chains [21]. The proteinase domain in epitheliasin contains eight Cys residues in conserved positions. By comparison with chymotrypsin, three of the Cys pairs (42/58, 168/182 and 191/220) that form disulfide bond loops around His⁵⁷, Met¹⁸⁰ and Ser¹⁹⁵ are conserved in epitheliasin. Although the other two cysteines (Cys¹²² and Cys¹³⁶) are located in conserved positions, their pairing counterparts Cys¹ and Cys²⁰¹ that are involved in interchain disulfide bonds are absent. This suggests that epitheliasin is likely distinct from enteropeptidase and other multidomain serine proteinases in that it lacks disulfide bond(s) between the proteinase motif and the rest of the protein [22]. Thus, the mechanism of association of the two chains in epitheliasin is not clear.

Three asparagine-linked glycosylation sites are present in epitheliasin, Asn¹¹¹ located at the beginning of the LDLRA domain of the protein, Asn²¹² located in the SRCR domain and Asn⁴⁷⁴ located in the proteinase domain (see Fig. 1). Other features of the deduced primary structure of the protein include a cAMP- or cGMP-dependent protein kinase phosphorylation site (Lys²⁴⁹-Ser²⁵²). Two protein kinase C phosphorylation sites are present in the cytoplasmic domain (Thr⁷⁷-Lys⁷⁹ and Thr⁸⁰-Lys⁸²), three in the SRCR domain (Ser¹⁶²-Arg¹⁶⁴, Ser²³¹-Arg²³³, Ser²³⁷-Arg²⁴⁹), one between the SRCR domain and the proteinase domain (Ser²⁵⁷-Lys²⁴⁹), and one in the proteinase domain (Thr⁴⁴⁵-Lys⁴⁴⁷). Three casein kinase II phosphorylation sites are present, two in the LDLRA domain (Ser¹¹³-Glu¹¹⁶, Ser¹¹⁶-Glu¹¹⁹), and the last one in the proteinase domain (Ser²⁶¹-Asp²⁶⁴). Finally, an ATP/GTP-binding site motif A is present in the proteinase domain of epitheliasin, from Ile³⁷⁹ to Ala³⁹⁶. This motif is found in a number of proteins including those in the myosin and Ras families. The relevance of these various sites in epitheliasin is not presently known.

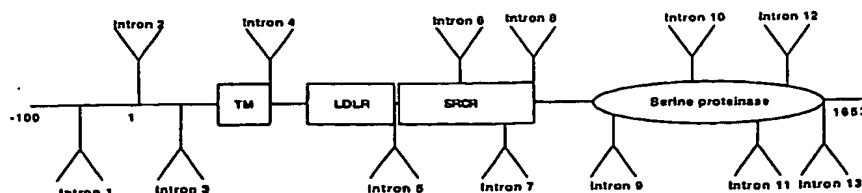


Fig. 3. Schematic representation of the genomic organization of epitheliasin. The intron placements are depicted in relationship to the domains of the mouse epitheliasin protein. The numbering represents nucleotides.

3.4. Genomic organization

The epitheliasin gene contains 14 exons separated by 13 introns (Fig. 3). The first exon is located in the 5' untranslated region. The last exon contains 9 bp of the coding sequence, the stop codon and the 3' untranslated region. The exon distribution reflects the organization of the deduced protein. Exon 2 and 3, respectively 68 and 220 nt (M^1-S^{78}), encode for the cytoplasmic domain. Exon 4, 87 nt, ($K^{79}-F^{107}$) encodes for the transmembrane domain. Exon 5, 117 nt, ($D^{109}-R^{146}$) encodes for the LDLR domain ($C^{112}-C^{147}$). An unusual feature of epitheliasin is that the SRCR domain is encoded by three exons, 6–8, respectively 130 nt, 111 and 44 nt ($C^{147}-I^{241}$). Usually SRCR domains are encoded by one or two exons, in regard to type B or type A, respectively. Exons 9–13, respectively 169, 176, 96, 143 and 153 nt, ($E^{242}-R^{490}$) encode for the serine protease domain. Vertebrate serine protease-like genes have been grouped into five classes based on intron positions [23]. The gene organization of the epitheliasin protease domain is typical of second group containing members of the trypsin family of serine proteases and consisting of five exons with each of the three components of the catalytic triad encoded by sequences in a different exon. In epitheliasin, the catalytic histidine is located in exon 9, the aspartic in exon 10 and the serine in exon 13. In general, the organization of epitheliasin is similar to that of other multiple domain serine proteinases. Each domain is coded in an independent manner by one or more exons. A common feature among all multi-domain protease cloned to date is the five exons coding for the serine proteinase domain [24].

As shown in Table 1, all intron/exon junctions contain the expected GT splice donor and AG splice acceptor sites and conform to the consensus sequences established for intronic donor and acceptor splice signals [25]. Four introns are inserted between codons (type 0 splice junction), five are after the first nucleotide in a codon (type I splice junction), and four after the second nucleotide codon (type II splice junction). Six bands were strongly positive by Southern analysis with sizes of 7000, 5000, 2700, 1400, 1200 and 900 nt. Adding the size of the fragments indicates that the epitheliasin gene is approximately 18 kb.

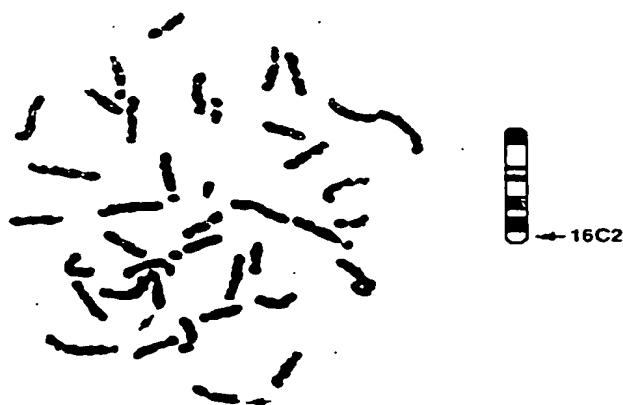


Fig. 4. In situ hybridization of a biotin-labeled epitheliasin probe to mouse metaphase cells. The chromosome 16 homologues are identified with arrows. Specific labeling was observed at chromosome band 16C2.

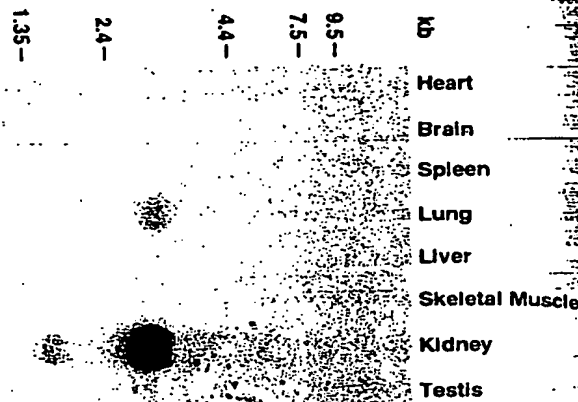


Fig. 5. Northern blot analysis of epitheliasin mRNA in various mouse tissues. Each lane contained 2 µg of poly(A)⁺RNA. The blot was hybridized to an epitheliasin cDNA probe.

3.5. Chromosomal assignment

FISH was performed on normal mouse chromosomes using a BAC containing the genomic sequence of epitheliasin (Fig. 4). These studies localized the epitheliasin gene to the telomeric region in the long arm of chromosome 16. The band localization was confirmed on G-banded chromosomes. The hybridization efficiency was 92.5%. No other serine proteinases have been localized to this region. The region is homologous with the so-called 'Down's syndrome region' of human chromosome region 21q22.2 and 21q22.3.

3.6. Expression of epitheliasin mRNA in vivo

The in vivo distribution of epitheliasin mRNA was investigated in adult mouse tissues by Northern blot analysis. As shown in Fig. 5, a prominent 2.8 kb transcript and a less prominent 1.5 kb transcript were observed in the kidney. Because of preliminary results that suggest an alternative polyadenylation site approximately 1.3 kb downstream from the initial polyadenylation site, we believe that the weaker signal actually represents the characterized cDNA. A prominent 2.8 kb signal was also seen in the lung and a weaker signal of similar size was observed in liver tissue. No signal was observed in heart, brain, spleen, testis or skeletal muscle. Of note, all tissues that express epitheliasin have epithelial cells as a prominent feature of their cellular makeup.

3.7. Immunohistochemical localization

Fig. 6A shows the kidney in which only tubular epithelial cells are stained with no staining of glomeruli. The staining is restricted to cells located in distal tubules. The staining is most intense at the apical pole of the cells, facing the lumen of the tubules. The staining is faint in the cytoplasm, basal and lateral side of the cells. Fig. 6B shows the lung in which staining is primarily limited to the apical surface of airway epithelial cells. Staining is minimal or absent in the vascular structure and alveolar spaces. No staining was observed in control slides. Further analysis by in situ hybridization using a 300 bp epitheliasin riboprobe demonstrated that the pattern of gene expression was the same as that of protein expression (data not shown).

Fig. 6. In situ hybridization of epitheliasin mRNA in bronchial and colonic tissue (data not shown).

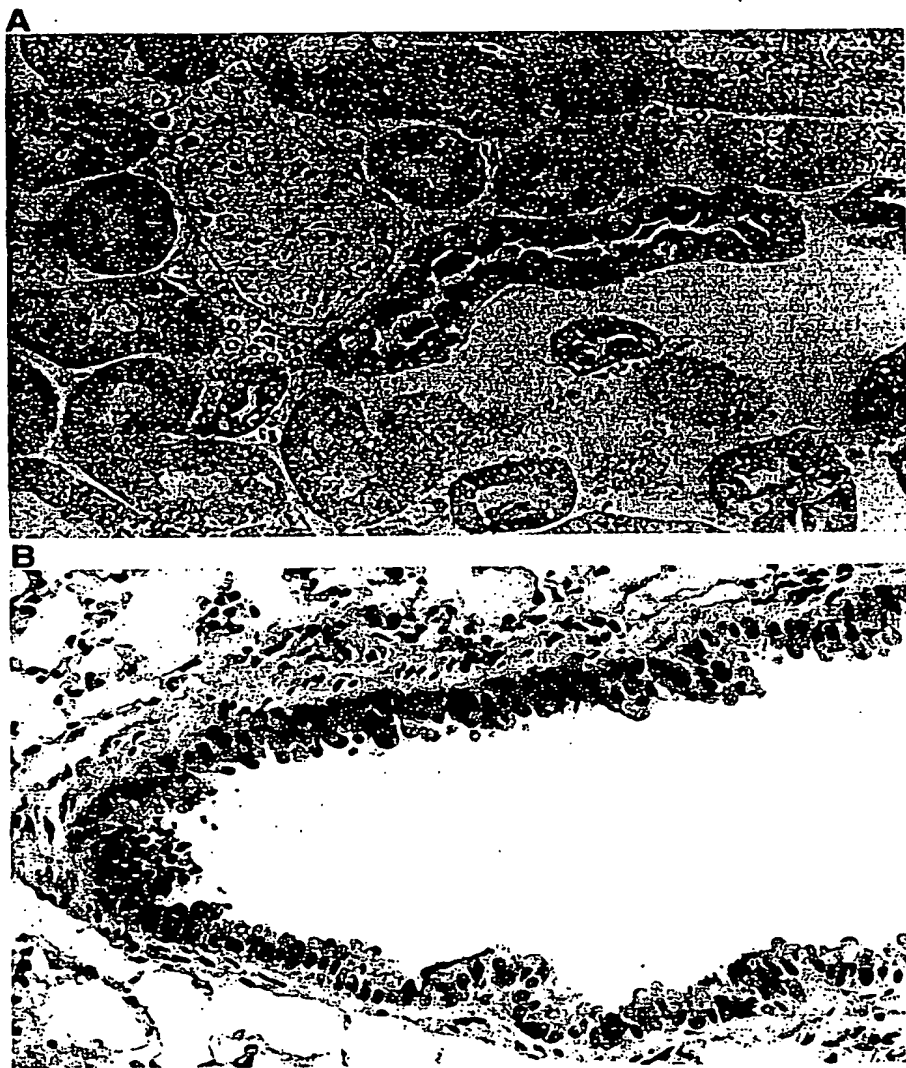


Fig. 6. Immunohistochemical localization of epitheliasin in adult mouse tissue. A: A section from the kidney (magnification 20 \times). Positive staining is seen in apical region of renal distal tubule epithelial cells. B: A section from lung (magnification 20 \times). Positive staining is seen in bronchial epithelial cells. No stain was observed in control sections in which normal rabbit serum substituted for rabbit anti-mouse epitheliasin (data not shown).

not shown). These results support the epithelial and membrane localization of epitheliasin.

During the course of this investigation Paolini-Giacobino and colleagues reported on a human cDNA cloned by exon trapping named TMPRSS2 [3]. The portion of the TMPRSS2 cDNA that was reported has approximately 80% sequence identity to epitheliasin. However, the tissue distribution of epitheliasin and TMPRSS2 is strikingly different. While epitheliasin is highly expressed in the mouse kidney, no expression of TMPRSS2 was observed in the human kidney. In contrast, no expression of epitheliasin was observed in the mouse heart or brain, while a high level of expression of

TMPPRSS2 was observed in human heart and an intermediate level in brain. Moreover, the size of epitheliasin of the mRNA transcript (2.8 kb) and that of TMPRSS2 (3.8 kb) are different. Whether TMPRSS2 is the human orthologue of epitheliasin or a closely related gene product will require further study.

The biological role of epitheliasin is not known. The homology with CAPI and apical membrane distribution raise the possibility that epitheliasin may activate ion transport channels of the plasma membrane. In addition, cell-surface proteinases of normal and malignant cells are thought to play roles in cell growth, chemotaxis, endocytosis, exocytosis,

blood coagulation, fibrinolysis and tissue invasion during metastasis [26]. While the function of the non-proteinase domains is unexplored, the presence of these domains with a modular organization represents a common feature of regulatory serine proteinases (e.g. proteinases of the fibrinolytic and blood coagulation systems). Studies of the kinetic effects of deleting the non-proteinase domain from enteropeptidase clearly implicate it in the recognition of macromolecular substrates and inhibitors [21].

Acknowledgements: The work was supported by HL 50153 and HL 37615 from the NHLBI. We gratefully acknowledge the assistance of Dr. Kurt Albertine and Zhengming Wang for the immunohistochemical studies. GenBank accession number for epitheliasin nucleotide sequence: Bank11243070 AF113596.

References

- [1] Baricos, W.H. and Shah, S.V. (1991) *Kidney Int.* 40, 161–173.
- [2] Vallet, V., Chraïbi, A., Gaeggeler, H.P., Horisberger, J.D. and Rossier, B.C. (1997) *Nature* 389, 607–610.
- [3] Paoloni-Giacobino, A., Chen, H., Peitsch, M.C., Rossier, C. and Antonarakis, S.E. (1997) *Genomics* 44, 309–320.
- [4] Pinkel, D., Straume, T. and Gray, J.W. (1986) *Proc. Natl. Acad. Sci. USA* 83, 2934–2938.
- [5] Kozak, M. (1986) *Cell* 44, 283–292.
- [6] von Heijne, G. and Manoil, C. (1990) *Protein Eng.* 4, 109–112.
- [7] High, S. (1992) *BioEssays* 14, 535–540.
- [8] Semenza, G. (1986) *Annu. Rev. Cell Biol.* 2, 255–313.
- [9] Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., Umeyama, H., Inoue, H., Tatabashi, T. and Takahashi, K. (1994) *J. Biol. Chem.* 269, 19976–19982.
- [10] Kazama, Y., Hamamoto, T., Foster, D.C. and Kiesel, W. (1995) *J. Biol. Chem.* 270, 66–72.
- [11] Yamaoka, K., Masuda, K.-I., Ogawa, H., Takagi, K.-I., Uemoto, N. and Yasuoka, S. (1998) *J. Biol. Chem.* 273, 11895–11901.
- [12] Sudhof, T.C., Goldstein, J.L., Brown, M.S. and Russell, D.W. (1985) *Science* 228, 815–822.
- [13] Resnick, D., Pearson, A. and Krieger, M. (1994) *Trends Biochem. Sci.* 19, 5–8.
- [14] Whitney, G.S., Starling, G.C., Bowen, M.A., Modrell, B., Siadak, A.W. and Aruffo, A. (1995) *J. Biol. Chem.* 270, 18187–18190.
- [15] Bowman, A. and Drummond, A.H. (1984) *Br. J. Pharmacol.* 81, 665–674.
- [16] Bresnahan, P.A., Hayflick, J.S., Molloy, S.S., and Thomas, G. (1993) in: *Mechanisms of Intracellular Trafficking and Processing of Proproteins* (Loh, Y.P., Ed.), pp. 225–250. CRC Press, Boca Raton, FL.
- [17] Creemers, J.W.M., Siezen, R.J., Roebroek, A.J.M., Ayoubi, T.A.Y., Huylebroeck, D. and Van de Ven, W.J.M. (1993) *J. Biol. Chem.* 268, 21826–21834.
- [18] Hatsuzawa, K., Nagahama, M., Takahashi, S., Takada, K., Murakami, K. and Nakayama, K. (1992) *J. Biol. Chem.* 267, 16094–16099.
- [19] Molloy, S.S., Bresnahan, P.A., Leppla, S.H., Klimpel, K.R. and Thomas, G. (1992) *J. Biol. Chem.* 267, 16396–16402.
- [20] Van de Ven, W.J.M. and Roebroek, A.J.M. (1993) *Crit. Rev. Oncog.* 4, 115–136.
- [21] Lu, D., Yuan, X. and Ninglong Z. S.J.E. (1997) *J. Biol. Chem.* 272 (50), 31293–31300.
- [22] Delabar, J.M., Tehophile, D., Rahmani, Z., Chettouh, Z., Blouin, J.L., Prieur, M., Noel, B. and Sinet, P.M. (1993) *Eur. J. Hum. Genet.* 1, 114–124.
- [23] Irwin, D.M., Robertson, K.A. and MacGillivray, R.T. (1988) *J. Mol. Biol.* 200, 31–45.
- [24] Cool, D.E. and MacGillivray, R.T. (1987) *J. Biol. Chem.* 262, 13662–13673.
- [25] Breathnach, R. and Chambon, P. (1981) *Annu. Rev. Biochem.* 50, 349–383.
- [26] Bond, J.S. (1991) *Biomed. Biochim. Acta* 50, 775–780.

Abstract
protein
protein
mitochondrial
proteins
magnetic
the con
TOM5
that is
forms a
structure
© 2001

Key wor
Nuclear

1. Intro

TOM
membran
protein
chondri
brane p
portion
TOM5
TOM22
proteins
quence
the incl
which p
brane p
for nat
linking
with pr
positive
id surfa
but littl
interact
conform
and att
lides.

*Corres
E-mail:

Also c

0014-5781
PII: S 0

Exhibit 17

Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains

(serine proteases/trypsinogen activation)

YASUNORI KITAMOTO*, XIN YUAN†, QINGYU WU*, DAVID W. MCCOURT†, AND J. EVAN SADLER*†‡

[†]Howard Hughes Medical Institute, *Departments of Medicine and Biochemistry and Molecular Biophysics, The Jewish Hospital of St. Louis, Washington University School of Medicine, St. Louis, MO 63110

Communicated by Earl W. Davie, April 19, 1994

ABSTRACT Enterokinase is a protease of the intestinal brush border that specifically cleaves the acidic propeptide from trypsinogen to yield active trypsin. This cleavage initiates a cascade of proteolytic reactions leading to the activation of many pancreatic zymogens. The full-length cDNA sequence for bovine enterokinase and partial cDNA sequence for human enterokinase were determined. The deduced amino acid sequences indicate that active two-chain enterokinase is derived from a single-chain precursor. Membrane association may be mediated by a potential signal-anchor sequence near the amino terminus. The amino terminus of bovine enterokinase also meets the known sequence requirements for protein N-myristoylation. The amino-terminal heavy chain contains domains that are homologous to segments of the low density lipoprotein receptor, complement components C1r and C1s, the macrophage scavenger receptor, and a recently described motif shared by the metalloprotease meprin and the *Xenopus* A5 neuronal recognition protein. The carboxyl-terminal light chain is homologous to the trypsin-like serine proteases. Thus, enterokinase is a mosaic protein with a complex evolutionary history. The amino acid sequence surrounding the amino terminus of the enterokinase light chain is ITPK-IVGG (human) or VSPK-IVGG (bovine), suggesting that single-chain enterokinase is activated by an unidentified trypsin-like protease that cleaves the indicated Lys-Ile bond. Therefore, enterokinase may not be the "first" enzyme of the intestinal digestive hydrolase cascade. The specificity of enterokinase for the DDDDK-I sequence of trypsinogen may be explained by complementary basic-amino acid residues clustered in potential S2–S5 subsites.

All animals need to digest exogenous macromolecules without destroying similar endogenous constituents. The regulation of digestive enzymes is, therefore, a fundamental requirement (1). Vertebrates have solved this problem, in part, by using a two-step enzymatic cascade to convert pancreatic zymogens to active enzymes in the lumen of the gut. The basic features of this cascade were described in 1899 by N. P. Schepovalnikov, working in the laboratory of I. P. Pavlov (2). Extracts of the proximal small intestine were shown to strikingly activate the latent hydrolytic enzymes in pancreatic fluid. Pavlov considered this intestinal factor to be an enzyme that activated other enzymes, or a "ferment of ferments," and named it "enterokinase." The importance of this protease cascade is emphasized by the life-threatening intestinal malabsorption that accompanies congenital deficiency of enterokinase (3, 4).

Enterokinase activates bovine trypsinogen by cleaving after the sequence VDDDDK, releasing an amino-terminal activation peptide (5, 6). The acidic DDDDK sequence of the trypsinogen-activation peptide is conserved among verte-

brates (7), except for the similar sequences of trypsinogens from lungfish (IEEDK and LEDDK) and African clawed frog (FDDDK). Enterokinase prefers substrates with the sequence DDDDK, whereas the presence of aspartate residues markedly inhibits the ability of trypsin to cleave such substrates (8). For example, toward bovine trypsinogen the catalytic efficiency of enterokinase is 12,000-fold (porcine) (9) or 34,000-fold (bovine) (10) greater than that of bovine trypsin. This reciprocal specificity protects trypsinogen against autoactivation by trypsin and promotes activation by enterokinase in the gut.

Enterokinase has been purified from porcine (11), bovine (10, 12, 13), human (14), and ostrich intestine (15). With the possible exception of human enterokinase, which was suggested to be a heterotrimer (14), enterokinase appears to be a disulfide-linked heterodimer with a heavy chain of 82–140 kDa and a light chain of 35–62 kDa. Mammalian enterokinases contain 30–50% carbohydrate, which may contribute to the apparent differences in polypeptide masses. The heavy chain is postulated to mediate association with the intestinal brush border membrane (16), although no direct evidence for this function has been reported. The light chain contains the catalytic center. Based on susceptibility to inhibition by chemical modification of the active-site serine and histidine residues (9–11, 17) and on the partial amino acid sequence (18) and cDNA sequence of the bovine enterokinase light chain (19), enterokinase is a member of the trypsin-like family of serine proteases.

Enterokinase stands at or near the top of a regulatory enzyme cascade that successfully limits the activity of digestive hydrolases to the gut, but there is no structural explanation for enterokinase membrane localization, substrate specificity, or expression specifically in the proximal small intestine. To address these questions we have characterized cDNA clones for bovine and human enterokinase.[§]

MATERIALS AND METHODS

Materials. Purified calf enterokinase (EK-3, 131 units/ μ g) was from Biozyme Laboratories (San Diego). Fresh bovine tissues were from a local abattoir.

Amino Acid Sequencing. Enterokinase (16 μ g) was reduced with 0.5% (vol/vol) 2-mercaptoethanol, separated by electrophoresis (20), transferred to an Immobilon P membrane (Millipore) by electroblotting, and stained with Coomassie brilliant blue. The excised light-chain band (\approx 47 kDa) was subjected to automated Edman degradation with an Applied Biosystems model 470A sequencer (21) equipped with a model 120A phenylthiohydantoin analyzer.

[‡]To whom reprint requests should be addressed at: Howard Hughes Medical Institute, Washington University, 660 South Euclid, Box 8022, St. Louis, MO 63110.

[§]The sequences reported in this paper have been deposited in the GenBank data base (accession nos. U09859 and U09860).

Isolation of cDNA Clones. RNA was extracted (22) from bovine duodenum and proximal small intestine. Single-stranded cDNA was prepared from total RNA (10 μ g) using avian myeloblastosis virus reverse transcriptase and an oligo(dT) primer (cDNA cycle kit, Invitrogen). The cDNA was used for PCR amplification (30 cycles of 2-min annealing at 58°C, 2-min extension at 72°C, and 1-min denaturation at 94°C) with sense primer 5'-TAY GAR GGI GCI TGG CCI TGG GT-3' and antisense primer 5'-AAT GGG ACC GCC IGA RTC ICC-3'. Products were analyzed by Southern blotting and hybridization with ³²P-labeled oligonucleotide probe 5'-STI WCI GCI GCC CAC TG-3'. The positive 572-bp product was cloned to yield pBEK1.

Additional clones were obtained by radiolabeling the cDNA insert of pBEK1 with [³²P]dCTP (23) and screening of bovine or human small intestine Agt11 cDNA libraries (Clontech) or by using oligonucleotides to screen 5' rapid amplification of cDNA ends (RACE) libraries (24). RACE libraries were constructed with the 5' RACE system (GIBCO/BRL) using bovine intestinal RNA and one of two sets of enterokinase-specific primers: set 1, 5'-TTA TTG TCT TCA TCA GAG CCA TC-3', 5'-TGG ACA GTT TAA TTC TCC ATC ACA-3', 5'-ATC AAT TGC TAT GTA CTT TAG AGC-3'; set 2, 5'-ATT GAG ACA TTT CCT GTG ATA TCA ATG CTG-3', 5'-TGT GGA AAG TGA CCA GTT GGC TGG ATT TAT-3', 5'-GCC TTG AAT CAG TTC TTC TT-3'. DNA sequences were determined on both strands (25).

DNA Sequence Analysis. Sequences were compared to GenBank and EMBL data bases at the National Center for Biotechnology Information using the BLAST network server (26). Sequence alignments and consensus sequences were prepared and analyzed with the programs PILEUP and GAP of the Genetics Computer Group (version 7.1, Madison, WI). The significance of GAP alignments was evaluated by comparing the optimal alignment score (x) to the mean (μ) and SD (σ) of scores obtained for 30 alignments of randomized sequences, using the normal distribution to estimate the probability that the alignment could occur by chance.

RESULTS AND DISCUSSION

Isolation of cDNA Clones. The bovine enterokinase light chain was reported to contain the motif YEGAWPWYV at residues 8–16 (18); the underlined residues are not conserved in other serine proteases. Thirty-one residues of the amino-terminal sequence of the bovine enterokinase light chain were determined, and the previously reported sequence was confirmed, except that arginine rather than tyrosine was identified at cycle 8. This sequence was used to design a degenerate 23-mer "sense" primer that would be relatively specific for enterokinase. A degenerate 21-mer "antisense" primer was based on the conserved GDSGGPL motif that contains the active-site serine of serine proteases. Upon PCR with a bovine small intestine single-stranded cDNA template, the major product hybridized to a probe based on the conserved sequence near the active-site histidine. The corresponding clone pBEK1 was used to isolate overlapping cDNAs from bovine and human small intestine cDNA libraries.

The composite cDNA sequence for bovine enterokinase spans 3923 nt. Beginning at nt 113 there is an ATG codon and open reading frame of 3105 nt, a stop codon plus 3' untranslated region of 643 nt, and a poly(A) tail of 63 nt. A polyadenylation signal of AATAAA is present 25 nt before the poly(A) tail. The open reading frame encodes a polypeptide of 1035 amino acids with a calculated mass of 114.9 kDa. The translated amino acid sequence after residue 800 (Fig. 1) was identical to the 31 residues determined by Edman degradation of the enterokinase light chain, confirming that the cDNA encodes enterokinase. A segment of 81 nt that encodes amino acid residues Ala-166–Pro-192 was present in three cDNA

clones but absent in one (Fig. 1). This sequence is not delimited by splice sites and therefore may be encoded by an exon that is occasionally absent due to alternative splicing. This segment also could represent a length polymorphism.

The partial cDNA sequence for human enterokinase corresponds to amino acids 765–1035 encoded by the bovine sequence. In the region of overlap, the open reading frames of the bovine and human nucleotide sequences are \approx 85% identical, and the encoded amino acid sequences are \approx 84% identical. The 3' untranslated regions are less conserved, exhibiting \approx 67% sequence identity over 572 nt.

By Northern blotting, an enterokinase mRNA species of \approx 4.4 kb was detected in human small intestine, but not in leukocytes, colon, ovary, testis, prostate, thymus, spleen, pancreas, kidney, skeletal muscle, liver, lung, placenta, brain, or heart (data not shown). This result is consistent with the studies of Pavlov on the distribution of enterokinase (2) and the immunohistochemical localization of enterokinase in the brush border of duodenum and jejunum (27).

Structure of the Enterokinase Catalytic Domain. In agreement with LaVallie *et al.* (19), amino acid residues 801–1035 correspond to the enterokinase light chain, which has a predicted mass of 26.3 kDa, compared with 47 kDa observed for purified bovine intestinal enterokinase (data not shown). The difference reflects glycosylation of the light chain. There are three and four potential N-linked glycosylation sites, respectively, in the bovine and human enterokinase light chains, and digestion of bovine enterokinase with peptide:N-glycosidase F reduces the apparent mass of the light chain from 47 kDa to 35 kDa (data not shown).

The enterokinase protease domain was compared with other serine proteases for characteristic disulfide bond patterns and sequence similarity. Enterokinase is most similar to a subfamily of two-chain serine proteases that share 10 conserved cysteine residues and in which the activation peptide remains attached to the protease domain by a disulfide bond. The archetype of this group is chymotrypsin. By analogy to chymotrypsin (28, 29) and related proteases for which the disulfide bonds have been determined directly, the most likely pairings in enterokinase are as follows: Cys-788–Cys-912, Cys-826–Cys-842, Cys-926–Cys-993, Cys-957–Cys-972, and Cys-983–Cys-1011. The first of these disulfide bonds joins the heavy chain and light chain.

The amino acid sequence of the enterokinase protease domain is strikingly similar to the blood coagulation proteases factor XI (30) and prekallikrein (31) and to hepsin, an unusual serine protease with a possible transmembrane domain near the amino terminus (32). Enterokinase exhibits the expected conservation of serine protease sequence motifs; in particular, the active-site residues can be identified as His-841, Asp-892, and Ser-987 (Fig. 1). Compared with factor XI, hepsin, and chymotrypsin, the human enterokinase light chain has 41%, 44%, and 35% identical amino acid residues. The percentages for the bovine enterokinase comparisons are similar. Enterokinase and factor XI appear to share two potential N-linked glycosylation sites, whereas hepsin has no N-linked glycosylation sites.

The specificity of enterokinase for cleavage after lysine is consistent with the presence of Asp-981 at the base, and Gly-1008 and Gly-1018 at the sides of the specificity pocket or S1 subsite that binds the substrate P1 residue (Fig. 1). The requirement for aspartate in the P2–P5 positions suggests that the surface of enterokinase should provide electrostatic complementarity to negatively charged side chains. Examination of the homologous three-dimensional structure of chymotrypsin suggests that several exposed surface loops of enterokinase (Fig. 1, segments a–d) might contact these substrate residues. Within these segments, there are a few positively charged residues that are present in both bovine and human enterokinase but absent from related proteases with different

specificity for the P2-P5 substrate residues. In particular, the RRRK (human) or KRRK (bovine) sequences between residues 886-889 (Fig. 1, segment b) may interact directly with the aspartate residues in enterokinase substrates.

The synthesis of enterokinase as a single-chain protein poses a conceptual problem because it indicates that "proenterokinase" itself must be activated by proteolytic cleavage. The responsible protease could act on proenterokinase intracellularly during biosynthesis or extracellularly. Although the reaction could be autocatalytic, the participation of a separate protease seems more likely. In that case, enterokinase would not be strictly at the top of the digestive hydrolase cascade but would be in the second position at best. The amino-terminal isoleucine of the enterokinase light chain is preceded by Ser-Pro-Lys (bovine) or Thr-Pro-Lys (human), suggesting that enterokinase is activated by a trypsin-like enzyme. The identity and location of the proenterokinase activator may indicate another level in the control of digestion.

Structural Motifs of the Enterokinase Heavy Chain. The nucleotide sequence around the codon for Met-1 is AAAATGG, and that for Met-20 is GTCATGT. Only the former sequence matches at both positions -3 and +4 the consensus sequence proposed for translation initiation in vertebrate mRNAs (33), suggesting that initiation at Met-1 is

more likely. There is no in-frame termination codon within the available 112 nt of putative 5' untranslated sequence, so it is possible that the initiation codon remains to be cloned. However, initiation at Met-1 predicts a bovine enterokinase heavy chain of 800 amino acids with a mass of 88.6 kDa (Fig. 1), and this is consistent with the ~763 amino acids and ~84 kDa estimated by compositional analysis of purified enterokinase (12). By SDS/gel electrophoresis, the apparent mass of the heavy chain was ~116 kDa, decreasing to ~82 kDa after removal of N-linked oligosaccharides with peptide:N-glycosidase F (data not shown). This decrease in mass is consistent with the reported carbohydrate composition of enterokinase (10, 12), and there are 17 potential N-linked glycosylation sites in the sequence of the heavy chain (two are concatenated) (Fig. 1).

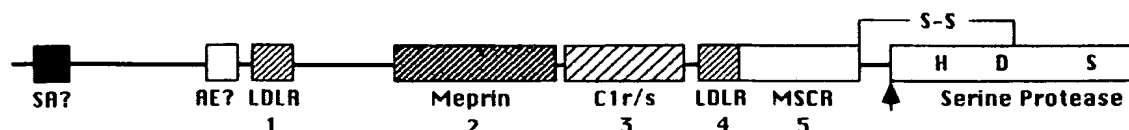
The hydrophobic 29-residue sequence from Val-19 through Val-47 could serve as a signal peptide. If it were not cleaved by signal peptidase, this segment could function as a signal-anchor sequence and account for the membrane association of enterokinase. The amino-terminal sequence also is compatible with the substrate specificity of myristoylCoA:protein N-myristoyltransferase (34), suggesting that Gly-2 may be myristoylated and thereby provide another mechanism for membrane targeting during biosynthesis.

The heavy chain of enterokinase contains five domains that are related to four different structural motifs found in other

EKbov	MGSKRVSFPR	HRSLLTYEVM	FAVLFPVILVA	LCAGLIAVSM	LSIQGSVKDA	AFGKSHEARG	TLKIISGATY	NPHLQDKLSV	DPKVLAFDIQ	QMIDDIFQSS	100
EKbov	NLKNEYKNSR	VLPQFNSII	VIPDLLFDQM	VSDKNVKEEL	IQGIEANRKS	QLVTFPHIDLN	SIDITASLEN	ESTISPATTS	EKLITSIPLA	TPGIVSIECP	200
EKbov	PDSRLCADAL	KYIAIDLPCD	GELNCPDGS	EDNATCATAC	DGRFLLTGSS	GSFEALHYPK	PSNYSAVCR	WIIRVNQGLS	IQLNFDYFNT	YYADVNIYE	300
EKbov	GMSSSKILRA	SLWSNNPGII	RIFSNQVTAT	FLIQSDSDY	IGPKVITYAF	NSKELANYEK	INCNPEDGFC	FWIQDLNDDN	EWERTQGSTF	PPSTGPTPDH	400
EKbov	TFGNSGFIYI	STPTGPGRRR	ERVGLLTPL	DPTPEQACLS	FWYMYGENV	YKLSIHLSD	QNMKTIFQK	EGNYQGNWNY	GQVTLNIVE	PKVSPYQFKN	500
EKbov	QILSDIALDD	ISLTYGICIV	SVYPEPTLVP	TPPPELPTDC	GGPHDLWEFL	TFTSINFPN	SYNQAFCIW	NLNAQKGNKI	QLHPQEPDLE	NIADVVEIRD	600
EKbov	GEGLDLSFLA	VYTGPGPVND	VFSTTNMTV	LPITDNMLAK	QGPKAHTFG	YGLGIPEPCK	EDNPQCKDGE	CIPLVNLCDG	FPCHKDGSDE	AHCVRLEPFT	700
EKhu	TDSSGLVQFR	IQSIWHVACA	ENMTQISDD	VQQLLGLGCT	NSVPTFTSG	GGPYVNLATA	PNASLILTPS	QCCLDSLIL	LQCNKSCGK	KLIV..TQEV	798
FXI	
Hepsin	
Chita	
Consensus	
EKhu	PKIVGGSNAK	EGAWPVVGL	YY...GGRLL	CGASLVSSDW	LVSAAHCYVG	RNLEPSKWTA	ILGLHMKSEL	TSPTVPRLI	DEIVINPHY.NRRRK	889
EKbov	PKIVGGSNSR	EGAWPVVGL	YF...DDQV	CGASLVSSDW	LVSAAHCYVG	RNLEPSKWTA	VLGLHMASL	TSPTIETRLI	DQIVINPHY.NRRRK	
FXI	PRIVGGTASV	RGEWPMQVTL	HTTSPTQRHL	CGSIIIGNQW	ILTAACHCFYG	..VESPKILR	VYSGILQDSE	IKEDTSFPGV	QELIINDQY.KMAES	
Hepsin	DRIVGGRTDS	LGRWPMQVSL	RY...DGAHL	CGSLLSGDW	VLTAACHCFPE	RNRVLSRWVR	FAG...AVAQ	ASPHGLQLGV	QAVVYHGGYL	PPRDPNSEEN	
Chita	SRIVNGEAV	PGSWPMQVSL	..QDKTGFHF	CGSLLINEHW	VVTAACHC...	..GVTTSDV	VVAGEFDQGS	SSEKIQLKI	AKVFNRSKY.NSLTI	
Consensus	PRIVGG-D--	-G-WPMQV-L	-Y---G-HL	CGSL-S-DW	VVTAACH-YG	RN-E-SKW--	V-G--M----	-SP---L-I	--IVIN--Y-	-----N----	
EKhu	DNDIAMHLE	FKV...TDYIQ	PICLPEENQV	PPPGRICSLA	GWGTVVY.CG	TTANILQEAD	VPLLSNERCQ	.QQMPEYHIT	ENMICAGYEE	GGIDSCQGDS	987
EKbov	NNDIAMHLE	MKV...TDYIQ	PICLPEENQV	PPPGRICSLA	GWGALIY.CG	STADVLQEAD	VPLLSNERCQ	.QQMPEYHIT	ENMVCAGYEA	GGVDSQCGDS	
FXI	GYDIALHLK	TTV...TDSQR	PICLPSKQDR	NVIYTDQMT	GWGRYKL.RD	KIQNTLQKAK	IPLVTNEECQ	.KRYRGHKIT	HMKICAGYRE	GGKDACQGDS	
Hepsin	SNDIALVHLS	SPLPLTEYIQ	FVCLPAQAQA	LVQDKICTVT	GWGNTQY.YG	QQAGVLQEAR	VPIISNDVCN	GADFYGNQIK	PMQFCAGYPE	GGIDACQGDS	
Chita	NNDITLLKLS	TAASFQSTVS	AVCLPSASDD	FAAGTTCTVT	GWGLTRYTNA	NTPDRLQOAS	LPLLSNTNCG	.KKYWGTKIK	DAMICAG..A	SGVSSCMGDS	
Consensus	-NDIAL-HLE	--VNYTDYIQ	PICLP---Q-	F--G-C--T	GWG---Y-G	-TA-VLQEA-	VPLLSNE-CQ	---Y-G--IT	E-MICAGY-E	GG-DSCQGDS	
EKhu	GGPLMCQEN.	...NRWFLAG	VTSFGYK.CA	LPNRPVYAR	VSRPTEWIS	FLH.....	
EKbov	GGPLMCQEN.	...NRWFLAG	VTSFGYK.CA	LPNRPVYAR	VSRPTEWIS	FLH.....	1035
FXI	GGPLSCQEN.	...EVMHLVG	ITSWEG.CA	QRERPGVYTN	VVEYVDWILE	KTQAV.....	
Hepsin	GGPFVCEDSI	SRTFPRWLCO	IVSWGTC.CA	LAQKPGVYTK	VSDPFEWIFQ	AIKTHSEASG	MVTQL.....	
Chita	GGPLVCKQEN	...GAWTLVG	IVSWGSSCS	.TSTPGVYAR	VTALVNVVQ	TLAAN.....	
Consensus	GGPL-C-EN-	---RW-L-G	ITSWG---CA	L--RPGVYAR	V--F-EWIO-	-L-----	

FIG. 1. Translated amino acid sequence of enterokinase cDNA clones and alignment with other serine proteases. The aligned sequences include human enterokinase (EKhu), bovine enterokinase (EKbov), human factor XI (FXI), human hepsin (Hepsin), bovine chymotrypsinogen A (Chita), and a consensus sequence. Numbering at right refers to the translated sequence of bovine enterokinase. Cysteine residues are in boldface type. Potential N-linked glycosylation sites are in boldface underlined type. The potential signal-anchor sequence is double underlined. The potential alternative exon is indicated by a dotted underline. Sequence motifs in the heavy chain are indicated by lettered underlines (a-d). The cleavage site for zymogen activation (Δ), active site residues (*), and residues in the specificity pocket or S1 subsite (\dagger) are indicated below the consensus sequence.

A



B

	1						52
EK1 (199-239)	C. PPDRLCA	D. ALKYIAID	LPCDGLNCP	DGSDENKTC	ATA.....		
EK4 (659-693)	C. KEDNFQCK	D. .GECIPLV	NLCDGFPCHK	DGSD. .AHC		
LDLR1 (6-46)	C. ERNEFQCC	D. .GKCISYK	WVCDGSAECQ	DGSDSQETC	LSVT.....		
LDLR2 (47-87)	C. KSGDFSCG	GRVNRICPQF	WRCDGQVDCD	NGSDE. .QGC	PPKT.....		
LDLR3 (88-126)	C. SQDFPRCH	D. .GKCISRQ	FVCDSDRCL	DGSD. .ASC	PVLT.....		
LDLR4 (127-175)	C. GPASFPCH	S. .STCIPQL	WACDNDPDC	DGSDWEPQRC	RGLYVFGQDS	SP	
LDLR5 (176-214)	C. SAFEFHCL	S. .GECIHSS	WRCDGDPCK	DKSD. .ENC	AVAT.....		
LDLR6 (215-254)	C. RPDEFQCS	D. .GNCIHGS	RQCDREYDCK	DMSDE. .VGC	VNVT.....		
LDLR7 (255-296)	CEGPNKFKCH	S. .GECITLD	KVCNMARDCR	DMSDEPIKEC	GTNE.....		
Consensus	C---EF-C-	D--G-CI--	W-CDG--DC-	DGSD-----	---T-----		

C

	1								90
EK2 (358-443)	YEKINCNF..	..EDGFCFWI	QDLNDDNEW	RTQGSTFPSP	TGPTFDHTFG	NESGFIYISTP	TGPGRRERV	GLLTLPDPT	PEQACLSFWY
A5xen (646-727)	HSDLDCKFGW	GSQKTVCMWQ	HDISSDLKWA	VLNSKTGP..VQD.H	TGDNFIYSE	ADERHEGRAA	RLMSPVSSS	RSACLTPFWY
MeprinA (276-360)	TLLDHCDFEK	..TNVCGMI	QGTDDADWA	H. GDSSSQEQ	VDHTLVEQ.C	KGAGYFMFFN	TSLGARGEAA	LLESRLYPK	RKQCCLOPFY
MeprinB (261-346)	SFMDSCDFEL	..ENICGMI	QSSQDSADWQ	RLSSQVLSGPE	NDHSHMGQ.C	KDSGPFMHFN	TSTGNGGITA	MLESRLYPK	RGFCQVEFYL
Consensus	---D-CDFE-	---NVCG-I	QD--DDADWA	RL--ST-PP-	-DHT-V-Q-C	K-SGPF--FN	TS-G-RGEAA	-L-SRVLYPK	R-QQCL-FWY

	91								179
EK2 (444-520)	YMGENVVKL	SINISSDQNM	EKT....IF	QKEGNYQNM	NYGQVTLNET	VEFKVSFYGF	...KNQILSD	IALLDISL..	..TYGICNV
A5xen (728-812)	HMDGSHVGT	SIKLKYEEM	DFDQTL...W	TVSGNQGDQW	KEARVVLHKT	MKQYQVIVEG	TVGKG.SAGG	IAVDDIIAN	HISPSQCR
MeprinA (361-445)	KMTGSPADRF	EVWVRDDNA	GKVRQLAKIQ	TPQGDSDHNM	KIAHVTLNNE	KKFRYVFLGT	KGDPQNSGG	IYLDITL..	..TETPCRA
MeprinB (347-430)	YNSGSGNGQL	NVYTRYTAG	HQDGVLTQR	EIRDIPTGSW	QLYVTLQVT	EKFRVVFEGV	.GGPGASSGG	LSIDDINL..	..SETCPH
Consensus	YM-GS-VG-L	S---R-D-N-	-KD--L--I-	T--GN-G-NW	K-A-VTLNET	-KFRVVF-G-	-GG-G-SSGG	IA-DDI-L--	---ET-CFA

D

	1								90
EK3 (540-619)	CGGPHDLWEP	NTTFTSINFP	NSYPNQAFCI	WNLNAQKGN	IQLHF.QEFD	LEN.....	.IADVVEIRD	GEGDSD.LFL	AVYTGPQPVN
Tolloid2 (468-550)	CGGDLKLTGD	QSI.DSPNYP	MDYMPDKECV	WRITAPDNHQ	VALKF.OSFE	LEK....HDG	CAYDFVEIRD	GNHSDS.RLI	GRFCGDKLPP
Tolloid3 (624-712)	CGGVVDATKS	NGSLYSPSY	DVYPNSKQCP	WEVVAAPNHA	VPLNF.SHPD	LECTRPHYTK	CNYDVLIYS	KMRDNLKKI	GIYCGHELPP
Tolloid4 (787-868)	CKFEI..TTS	YGVLSQSNYP	EDYPRNIYCY	WHFQVTLGHR	QLTTF.HDPE	VES....HQE	CIVDYVAIYD	GRSENS.STL	GIYCGGREPY
C1r1 (18-95)	..SIPIPQKL	FGEVTSPLFP	KYPNNFETT	TVITVPTGIR	VKLVF.QQFD	LEPS....EG	CFYDYVKISA	DKKS.....	GRFCQQLGSP
C1r2 (193-274)	CSSELY.TEA	SGYISSLEYP	RSYPPDLRCN	YSIRVERGLT	LHLKFLPEPD	IDDHQQVH..	CPYDQLQIYA	NGKN.....	I.GEFCQKORPP
Consensus	CGGE---TK-	-G-L-SPNYP	--YPN---CV	W-I-AP-AGH	V-L-F-Q-FD	LE-----H--	C-YDYV-IYD	G--D-S----	GI-CG---PP

	91								130
EK3 (620-651)DV	FSTINRMIVL	FITDNLAKQ	GFKANPTTGY					
Tolloid2 (551-581)NI	KTRSNQMYIR	FVSDSSVQKL	GFSALMLD.					
Tolloid3 (713-743)VV	NSEQSILRL	FYSDRTVQRS	GFVAKFVID.					
Tolloid4 (869-899)AV	IASTNEMPMV	LATDAGLQK	GFKATPVSE.					
C1r1 (96-135)	LGNPFGKKEF	MSQGNKMLT	FHTDPSNEEN	GTIMFYKGFL					
C1r2 (275-306)DL	DTSSNAVDLL	FPTDESQDSR	GWKLRYTTEI					
Consensus	-----DV	-S--N-M-L-	F-TD-S-Q--	GFKA-F----					

E

	1								90
EK5 (694-782)	VRLPNGTTDS	SGLVQFRIQS	IWHVACAENW	TTQISDDVQC	LLGLGTGNSS	VPTFTSGGGP	YVNLNTAPNG	SLILTFSSQC	LE.DSLILLQ
MSCR (349-428)	VRLVGGSGPH	EGRVEILHSG	QMGITCDDRW	EVVRQGVVCR	SLGYPGV...	..QAVHKAH	F.GGGTGP..	..IWLNEVFC	FGRE.SSIEE
Speract1 (43-121)	IRLIHGRTEN	EGSVEIYHAT	RWGGVCDMMW	HMENANVTCK	QLGFPGARQ.	...FYRRAY	P.GAHVTT..	..FWVYKMC	LGNE.TRLD
Speract2 (153-232)	LRMLGDPVN	EGTLETFWDC	AWGSCVCHTF	GTDPGNVACR	QMGYSRGVK.	...SIKTGPH	F.GFTSTGP..	..IILDAVDC	EGTE.AHTE
Speract3 (264-344)	IRLMDGSGPH	EGRVEIWHDD	AWGTICDDCW	DMADANVACR	QAGYRGAVK.	..ASGPKGED	F.GFTWAP..	..IHTSFVMC	TGVE.DRLID
Speract4 (382-464)	VRIV.GMGQG	QGRVEVSLGN	GWGRVCDPDW	SDHEARTVCY	HAGYKMGASR	AAGSAEVSAP	F.DLE.AP..	..FIIDGITC	SGVENETLSQ
Consensus	VRL--G-GP-	EGRVEI-H--	-WG-VCDW-W	---DANVCR	QLGY-GG---	---S-----A-	F-G--TAP--	---I----V-C	LG-E---L--

	91								116
EK5 (783-787)	CNYKS.....
MSCR (429-451)	CKIR.QWGR	AC..SHSEDA	GVTCTL						
Speract1 (122-145)	CYHRPYGRPW	LC..NAQWAA	GVECLP						
Speract2 (233-258)	CNMPVTPYQH	ACPYTHNDV	GVVCKP						
Speract3 (345-367)	CILR.DGWT	SC..YHVEDA	SVVCAT						
Speract4 (465-486)	CQMKV.SADM	TC...ATGTV	GVVCEG						
Consensus	C-MR-----	-C--H--DA	GVVCC--						

FIG. 2. Structural motifs in enterokinase. Numbers in parentheses refer to the amino acid residues represented in each aligned sequence. Bovine enterokinase (EK) residues are numbered as in Fig. 1. (A) Schematic structure of enterokinase, indicating the proposed signal-anchor sequence (SA), alternative exon (AE), numbered heavy chain domains (LDLR, low-density-lipoprotein receptor; MSCR, macrophage scavenger receptor), and serine protease domain with active-site residues histidine (H), aspartate (D), and serine (S). The cleavage site between the heavy and light chains (arrowhead) and disulfide bond connecting them are shown. (B) Alignment of EK domains 1 and 4 with cysteine-rich motifs of the LDL receptor (LDLR) (35). (C) Alignment of EK domain 2 with segments of *Xenopus laevis* A5 antigen (A5xen) (36), mouse meprin A (37), and rat meprin B (38). (D) Alignment of EK domain 3 with selected C1r/s-like domains of *Drosophila melanogaster* tollid (39), and complement component C1r (40). (E) Alignment of EK domain 5 with repeated domains of the mouse macrophage scavenger receptor type 1 (MSCR) (41) and the speract crosslinking protein from sea urchin sperm (42). The significance of alignments was estimated as described under *Materials and Methods*: EK1 or EK4 versus LDLR motifs, $P < 10^{-14}$; EK2 versus meprin motifs, $P < 10^{-23}$; EK3 versus C1r/s motifs, $P < 10^{-23}$; EK5 versus MSCR motifs, $P \approx 3.7 \times 10^{-5}$.

protein families, indicating that enterokinase is a mosaic protein with a complex evolutionary history. The particular combination of motifs is specific and surprising (Fig. 2A). Enterokinase domains 1 and 4 are homologous to an ≈ 40 -amino acid cysteine-rich repeat found in the amino-terminal domain of the low-density lipoprotein receptor and also in several complement proteins (Fig. 2B) (35).

Enterokinase domain 2 (Fig. 2C) is homologous to ≈ 170 -amino acid segments of meprins A and B, which are membrane-bound metalloproteases of renal glomeruli (37, 38). This domain also is homologous to a segment of the A5 protein of *X. laevis* (36), which may mediate neuronal recognition. For this structural motif, identified in four distinct vertebrate proteins, we propose the name "meprin domain."

Enterokinase domain 3 (Fig. 2D) is homologous to a family of ≈ 120 -amino acid repeats reported in complement serine protease C1r (40) and subsequently found in many proteins including the product of the *Drosophila* dorsal-ventral patterning gene *tolloid* (39). Interestingly, *tolloid* also encodes a separate metalloprotease domain that is homologous to the metalloprotease domains of meprins A and B.

Enterokinase domain 5 (Fig. 2E) is homologous to ≈ 110 -amino acid cysteine-rich motifs that are found in the macrophage scavenger receptor (41), the sea urchin spermatozoa speract receptor (42), and several lymphocyte cell-surface antigens (41). This domain in enterokinase is truncated at the carboxyl end.

The structural domains of the enterokinase heavy chain are found in proteins of the complement cascade, in endocytic receptors for diverse ligands including lipoproteins, in proteins that regulate development, in receptors that contribute to the specificity of egg fertilization, and in proteins of unknown function. The particular combination of structural motifs observed in the enterokinase heavy chain is unprecedented. The presence of potential ligand-binding domains suggests that interaction with other macromolecules, either in the cell membrane or in the lumen of the gut, might modulate enterokinase activation, substrate specificity, or inhibition.

For nearly a century enterokinase has been known as the principal activator of digestive hydrolases, and the same basic regulatory mechanism appears to be conserved among all vertebrates. The physiologic importance of this mechanism is emphasized by the severe malabsorption that accompanies human enterokinase deficiency (3, 4). The apparent requirement for proteolytic activation of proenterokinase suggests that yet another protease is required for the normal regulation of pancreatic zymogens. The isolation of cDNA clones for human and bovine enterokinase provides the means to address the regulation and structure-function relationships of this ancient, essential protease.

We thank Dr. Anja Schweizer and Dr. Jack Rorer (Washington University) for translations of articles from the original German, Dr. Heidi Hope (Washington University) for assistance with protein blotting, Lisa Westfield (Howard Hughes Medical Institute) for synthesis of oligonucleotides, and Cecil Buchanan (Washington University) for assistance in obtaining fresh bovine tissues. We also thank Prof. Tatsuo Sato (Kumamoto University, Kumamoto, Japan) for his encouragement and support. This research was supported, in part, by National Institutes of Health Grant HL14147 (Specialized Center of Research in Thrombosis).

1. Neurath, H. (1984) *Science* **224**, 350–357.
2. Pavlov, I. P. (1902) *The Work of the Digestive Glands* (Charles Griffin, London), 1st Ed.
3. Hadorn, B., Tarlow, M. J., Lloyd, J. K. & Wolff, O. H. (1969) *Lancet* **i**, 812–813.

4. Haworth, J. C., Gourley, B., Hadorn, B. & Sumida, C. (1971) *J. Pediatr.* **78**, 481–490.
5. Davie, E. W. & Neurath, H. (1955) *J. Biol. Chem.* **212**, 515–529.
6. Yamashina, I. (1956) *Acta Chem. Scand.* **10**, 739–743.
7. Brictaux-Gregoire, S., Schyns, R. & Florkin, M. (1972) *Comp. Biochem. Physiol. B* **42**, 23–39.
8. Abita, J. P., Delaage, M., Lazdunski, M. & Savrda, J. (1969) *Eur. J. Biochem.* **8**, 314–324.
9. Maroux, S., Baratti, J. & Desnuelle, P. (1971) *J. Biol. Chem.* **246**, 5031–5039.
10. Anderson, L. E., Walsh, K. A. & Neurath, H. (1977) *Biochemistry* **16**, 3354–3360.
11. Baratti, J., Maroux, S., Louvard, D. & Desnuelle, P. (1973) *Biochim. Biophys. Acta* **315**, 147–161.
12. Liepnieks, J. J. & Light, A. (1979) *J. Biol. Chem.* **254**, 1677–1683.
13. Fonseca, P. & Light, P. (1983) *J. Biol. Chem.* **258**, 14516–14520.
14. Magee, A. I., Grant, D. A. W. & Hermon-Taylor, J. (1981) *Clin. Chim. Acta* **115**, 241–254.
15. Naude, R. J., Da Silva, D., Edge, W. & Oelofsen, W. (1993) *Comp. Biochem. Physiol. B* **105**, 591–595.
16. Fonseca, P. & Light, A. (1983) *J. Biol. Chem.* **258**, 3069–3074.
17. Baratti, J. & Maroux, S. (1976) *Biochim. Biophys. Acta* **452**, 488–496.
18. Light, A. & Janska, H. (1991) *J. Protein Chem.* **10**, 475–480.
19. LaVallie, E. R., Rehemtulla, A., Racie, L. A., DiBlasio, E. A., Ferenz, C., Grant, K. L., Light, A. & McCoy, J. M. (1993) *J. Biol. Chem.* **268**, 23311–23317.
20. Laemmli, U. K. (1970) *Nature (London)* **227**, 680–685.
21. Hewick, R. M., Hunkapillar, M. W., Hood, L. E. & Dreyer, W. J. (1981) *J. Biol. Chem.* **256**, 7990–7997.
22. Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* **162**, 156–159.
23. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
24. Frohman, M. A., Dush, M. K. & Martin, G. R. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 8998–9002.
25. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
26. Gish, W. & States, D. J. (1993) *Nature Genet.* **3**, 266–272.
27. Hermon-Taylor, J., Perrin, J., Grant, D. A. W., Appleyard, A. & Magee, A. I. (1977) *Gut* **18**, 259–265.
28. Hartley, B. S. & Kauffman, D. L. (1966) *Biochem. J.* **101**, 229–231.
29. Brown, J. R. & Hartley, B. S. (1966) *Biochem. J.* **101**, 214–228.
30. Fujikawa, K., Chung, D. W., Hendrickson, L. E. & Davie, E. W. (1986) *Biochemistry* **25**, 2417–2424.
31. Chung, D. W., Fujikawa, K., McMullen, B. A. & Davie, E. W. (1986) *Biochemistry* **25**, 2410–2417.
32. Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K. & Davie, E. W. (1988) *Biochemistry* **27**, 1067–1074.
33. Kozak, M. (1991) *J. Cell Biol.* **115**, 887–903.
34. Rudnick, D. A., McWherter, C. A., Gokel, G. W. & Gordon, J. I. (1993) *Adv. Enzymol.* **67**, 375–430.
35. Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228**, 815–822.
36. Takagi, S., Hirata, T., Agata, K., Mochii, M., Eguchi, G. & Fujisawa, H. (1991) *Neuron* **7**, 295–307.
37. Jiang, W., Gorbea, C. M., Flannery, A. V., Beynon, R. J., Grant, G. A. & Bond, J. S. (1992) *J. Biol. Chem.* **267**, 9185–9193.
38. Johnson, G. D. & Hersh, L. B. (1992) *J. Biol. Chem.* **267**, 13505–13512.
39. Shimell, M. J., Ferguson, E. L., Childs, S. R. & O'Connor, M. B. (1991) *Cell* **67**, 469–481.
40. Leytus, S. P., Kurachi, K., Sakariassen, K. S. & Davie, E. W. (1986) *Biochemistry* **25**, 4855–4863.
41. Freeman, M., Ashenas, J., Rees, D. J. G., Kingsley, D. M., Copeland, N. G., Jenkins, N. A. & Krieger, M. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 8810–8814.
42. Dangott, L. J., Jordan, J. E., Bellet, R. A. & Garbers, D. L. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 2128–2132.

Exhibit 18

cDNA Sequence and Chromosomal Localization of Human Enterokinase, the Proteolytic Activator of Trypsinogen^{†,‡}

Yasunori Kitamoto,^{*,§} Rosalie Ann Veile,^{||} Helen Donis-Keller,^{||} and J. Evan Sadler^{*,†,‡}

Howard Hughes Medical Institute, Division of Human Molecular Genetics, Department of Surgery, Division of Hematology-Oncology, Department of Medicine, and Department of Biochemistry & Molecular Biophysics, The Jewish Hospital of St. Louis, Washington University School of Medicine, St. Louis, Missouri 63110

Received January 4, 1995[®]

ABSTRACT: Enterokinase is a serine protease of the duodenal brush border membrane that cleaves trypsinogen and produces active trypsin, thereby leading to the activation of many pancreatic digestive enzymes. Overlapping cDNA clones that encode the complete human enterokinase amino acid sequence were isolated from a human intestine cDNA library. Starting from the first ATG codon, the composite 3696 nt cDNA sequence contains an open reading frame of 3057 nt that encodes a 784 amino acid heavy chain followed by a 235 amino acid light chain; the two chains are linked by at least one disulfide bond. The heavy chain contains a potential N-terminal myristoylation site, a potential signal anchor sequence near the amino terminus, and six structural motifs that are found in otherwise unrelated proteins. These domains resemble motifs of the LDL receptor (two copies), complement component C1r (two copies), the metalloprotease meprin (one copy), and the macrophage scavenger receptor (one copy). The enterokinase light chain is homologous to the trypsin-like serine proteinases. These structural features are conserved among human, bovine, and porcine enterokinase. By Northern blotting, a 4.4 kb enterokinase mRNA was detected only in small intestine. The enterokinase gene was localized to human chromosome 21q21 by fluorescence *in situ* hybridization.

Enterokinase was discovered by N. P. Schepovalnikov, in I. P. Pavlov's laboratory, as an activity of small intestinal mucosa that dramatically increased the proteolytic activity of pancreatic fluid (Pavlov, 1902). Enterokinase later was shown to be an enzyme (Kunitz, 1939) that cleaves the amino-terminal activation peptide from trypsinogen to produce trypsin (Davie & Neurath, 1955; Yamashina, 1956). This reaction permits the subsequent activation of other pancreatic zymogens by trypsin. The physiologic importance of this two-step proteolytic cascade is indicated by the intestinal malabsorption that is caused by congenital deficiency of enterokinase (Hadorn et al., 1969; Haworth et al., 1971).

Enterokinase has been purified from bovine (Anderson et al., 1977; Liepnieks & Light, 1979; Fonseca & Light, 1983), porcine (Baratti et al., 1973), human (Magee et al., 1981), and ostrich intestine (Naude et al., 1993). In most preparations, enterokinase appears to be a disulfide-linked heterodimer composed of an 82–140 kDa heavy chain and a 35–62 kDa light chain, although a trimeric structure also has been proposed for human (Magee et al., 1981) and

porcine (Matsushima et al., 1994) enterokinase. Both chains of mammalian enterokinases contain 30–50% carbohydrate.

Recently, the full-length amino acid sequences of bovine (LaVallie et al., 1993; Kitamoto et al., 1994) and porcine (Matsushima et al., 1994) enterokinase and a partial sequence of human enterokinase (Kitamoto et al., 1994) were determined indirectly by cDNA cloning. Active enterokinase appears to be a two-chain protein derived from a single-chain precursor. The putative heavy chain contains a hydrophobic potential signal-anchor sequence near the amino terminus, as well as several domains that are homologous to structural motifs found in other proteins. The light chain contains the catalytic center, and enterokinase is a member of the trypsin-like family of serine proteases.

Many facts remain unknown concerning the structure and function of enterokinase. Although enterokinase appears to be an intrinsic membrane protein, the mechanism of membrane association is unknown. Furthermore, single-chain proenterokinase is proteolytically cleaved to generate active two-chain enterokinase, but the enzyme that is responsible for proenterokinase activation has not been identified.

To facilitate the study of human enterokinase membrane localization and zymogen activation, we have characterized cDNA clones that encode the complete amino acid sequence of human proenterokinase. These clones were employed to localize the human enterokinase gene to human chromosome 21q21 by fluorescence *in situ* hybridization.

EXPERIMENTAL PROCEDURES

Isolation of cDNA Clones. The partial human enterokinase cDNA insert contained in plasmid pHEK6 (Kitamoto et al., 1994) was labeled with [³²P]dCTP by a random primer method (Feinberg & Vogelstein, 1983) and employed to

[†] Supported in part by National Institutes of Health Grants HL14147 (J.E.S., Y.K.) and HG00469 (R.A.V., H.D.K.).

[‡] The DNA sequence (Figure 2) was deposited in the GenBank database under Accession Number U09860.

[§] Address correspondence to this author at the Howard Hughes Medical Institute, 660 South Euclid Ave., Box 8022, St. Louis, MO 63110.

^{||} Division of Hematology-Oncology, Department of Medicine.

[¶] Present address: Third Department of Internal Medicine, Kumamoto University School of Medicine, 1-1-1 Honjo, Kumamoto 860, Japan.

^{||} Division of Human Molecular Genetics, Department of Surgery.

[‡] Howard Hughes Medical Institute and Department of Biochemistry and Molecular Biophysics, The Jewish Hospital of St. Louis.

[®] Abstract published in *Advance ACS Abstracts*, April 1, 1995.

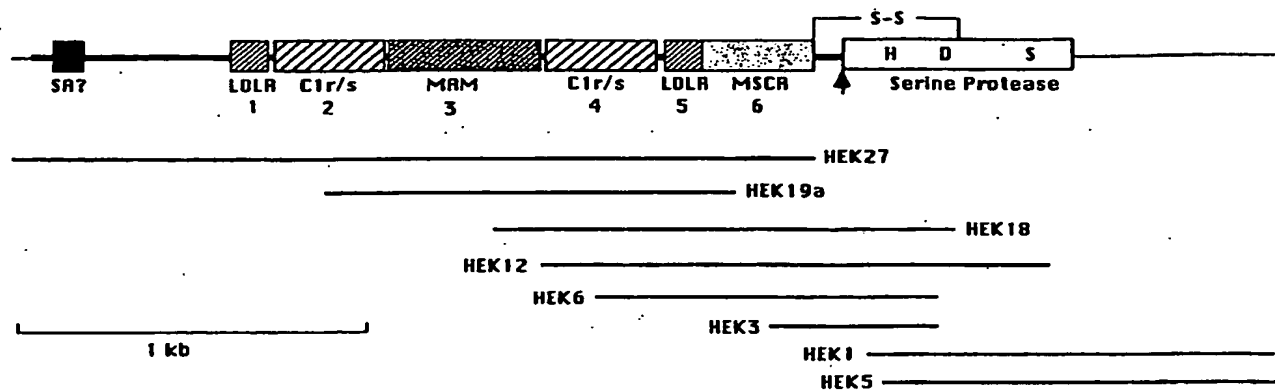


FIGURE 1: Domain structure of human enterokinase and map of enterokinase cDNA clones. The structure of the enterokinase cDNA is indicated schematically at the top. The 5' and 3' untranslated regions are indicated by thin lines (—) at the extreme left and right ends. The locations are indicated for a proposed signal-anchor domain (SA) and serine protease domain with active site histidine (H), aspartate (D), and serine (S) residues. The locations are shown of the cleavage site between the heavy and light chains (arrowhead) and of the predicted disulfide bond that connects them. The enterokinase heavy chain contains repeated motifs (numbered 1–6) that are homologous to domains of other proteins: LDLR, a low-density lipoprotein receptor cysteine-rich repeat (Sudhof et al., 1985); C1r/s, a repeat type found in complement components C1r and C1s (Leytus et al., 1986) and also found in the *Drosophila* dorsal-ventral patterning gene *tolloid* (Shimell et al., 1991); MAM, a domain homologous to members of a family defined by motifs in the mammalian metalloprotease meprin, the *X. laevis* neuronal protein A5, and the protein tyrosine phosphatase μ (Beckmann & Bork, 1993); MSCR, macrophage scavenger receptor cysteine-rich motif (Freeman et al., 1990) also found in sea urchin spermatozoa speract receptor (Dangott et al., 1989). The relationships among eight overlapping cDNA clones are indicated. The scale in kilobases (kb) of DNA is indicated at the bottom left.

screen a human small intestine cDNA library in the bacteriophage λ gt11 vector (Clontech). The cDNA inserts of plaque-purified isolates were subcloned into plasmid pBlue-script M13+ or pBluescript II KS+ (Stratagene) for DNA sequencing (Sanger et al., 1977).

DNA Sequence Analysis. Sequences were compared to GenBank and EMBL data bases at the National Center for Biotechnology Information using the BLAST network server (Gish & States, 1993). Sequence alignments and consensus sequences were prepared and analyzed with the programs pileup, gap, and pretty of the Genetics Computer Group (version 7.1, Madison, WI) as described previously (Kitamoto et al., 1994).

Northern Blotting. The insert of human enterokinase cDNA clone HEK1 or human β -actin (Gunning et al., 1983) was labeled with [32 P]dCTP (Feinberg & Vogelstein, 1983). A Northern blot of poly(A)+ RNA (Clontech) from assorted human tissues (2 μ g/lane) was hybridized (Sambrook et al., 1989) with the radiolabeled HEK1 insert (1 \times 10⁷ cpm/mL) and washed three times for 15 min at room temperature in 2 \times SSC¹ and 0.05% SDS (1 \times SSC is 15 mM sodium citrate, pH 7.0, 0.15 M NaCl). The final stringent wash condition was 50 $^{\circ}$ C, 15 min, in 0.1 \times SSC and 0.1% SDS. The blot was exposed to X-ray film for 10 days. The blot was stripped of radiolabeled HEK1 by immersion in 0.5% SDS for 10 min at 100 $^{\circ}$ C. The stripped blot was hybridized with the radiolabeled β -actin probe, washed as described above, and exposed to X-ray film for 2 h.

Gene Mapping by in Situ Hybridization. Fluorescence in situ hybridization was performed as described (Lichter et al., 1988). Human prometaphase chromosome spreads were prepared from cultured phytohemagglutinin-stimulated peripheral blood leukocytes from a male with a normal karyotype (46XY). Extended chromosomes were produced

by colchicine treatment (Yunis, 1976). Plasmids pHEK1 and pHEK6 contain the human enterokinase cDNA inserts of bacteriophage λ gt11 isolates HEK1 and HEK6, respectively, cloned into plasmid pBluescript M13+. Equal amounts were mixed of pHEK1 and pHEK6, and \approx 150 ng of DNA was labeled with biotin-11-dUTP by nick translation (Rigby et al., 1977). The biotinylated product was hybridized to human chromosomal spreads (Lichter et al., 1988). To detect sites of hybridization, slides were incubated sequentially with fluorescein isothiocyanate-conjugated avidin DCS (5 μ g/mL) and fluorescein isothiocyanate-conjugated goat anti-avidin D antibodies (5 μ g/mL), followed by counterstaining with 4,6-diamino-2-phenylindole dihydrochloride (200 ng/mL) and propidium iodide (200 ng/mL). After fluorescent hybridization, cytogenetic banding patterns were visualized by staining with Giemsa.

RESULTS AND DISCUSSION

Isolation of cDNA Clones. A human small intestine λ gt11 cDNA library was screened with the insert of a partial human enterokinase cDNA clone, HEK6 (Kitamoto et al., 1994). Seven positives were identified among 1.5 \times 10⁶ plaques screened. Clones HEK12, HEK18, and HEK19a were characterized further by restriction mapping and sequencing (Figure 1). The cDNA insert of HEK19a was employed to rescreen the library, and the longest clone obtained (HEK27) was sequenced.

The composite cDNA sequence of human enterokinase (Figure 2) was determined on both strands. Beginning at nt 41 there is an ATG codon and open reading frame of 3057 nt, followed by a stop codon and 3' noncoding region of 599 nt. The open reading frame encodes a polypeptide of 1019 amino acids with a calculated mass of 112.9 kDa. The coding regions of the human and bovine (Kitamoto et al., 1994) nucleotide sequences are \approx 85% identical, and the encoded amino acid sequences are \approx 82% identical. The 3' noncoding regions are less conserved, with \approx 67% identity between human and bovine enterokinase cDNA sequences

¹ Abbreviations: kb, kilobase; nt, nucleotide; SSC, standard saline citrate (15 mM sodium citrate, pH 7.0, 0.15 M NaCl); SDS, sodium dodecyl sulfate.

ACCAGACAGT	TCTTAAATTA	GCAAGCCTTC	AAAACCAAAA	ATGGCGTTCG	AAAGAGCCAT	ATCTTCTAGG	CATCATTCTC	TCAGCTCCTA	TGAATCATG	100
TTTGACAGCT	TCTTTOCCAT	ATTGGTAGTG	CTCTGTGCTG	GATTAATTCG	AGTATCTGCG	CTGACAATCA	AGGAATCCCA	ACGAGGTGCA	GCACCTGGAC	200
AGAGTCATGA	AGCCAGAGCG	ACATTTAAAA	TAACATCCGG	AGTTACATAT	AATCTTAATT	TGCAAGACAA	ACTCTCAGTG	GATTTCAAAG	TTCTTGCTTT	300
QSHHE	ARA	ATFK	ITS	GT	NP	LDK	LSV	DFK	VLA	87
TCAGCTTCAG	CAATGTATAG	ATGAGATCTT	TCTATCAAGC	AACTCGAAGA	ATGAATATAA	GAATCAAGA	GTTTTACAAT	TTGAAAATG	CAGCATTATA	400
DLO	QMI	DEIF	LS	NLK	NEY	NSR	V L Q	F E N G	S I I	120
GTGCTATTG	ACCTTTTCTT	TGCCAGTGG	GTGTCAGATC	AAAATGTAAA	AGAAGAAGCT	ATTCAAGGCC	TTGAAGCAAA	TAAATCCAGC	CAACTGGTCA	500
VVF	DLFF	AQW	VSD	QNV	EEL	I Q G	L E A N	K S S	Q L V	153
CTTCCATAT	TGATTGTAAC	AGCGTTGATA	TCTTAGACAA	GCTAACAACT	ACCACTATC	TGGCAACTCC	AGGAAATGTC	TCAATAGAGT	GCCTCGCTGG	600
T F H I	D L N	S V D	I L D K	L T T	T S H	L A T P	G N V	S I E	C L P C G	187
TTCAAGTCTT	TGACTGATG	CTCTAACGTG	TATAAAAGCT	GATTTATTTT	GTGATGGAGA	AGTAAACTGT	CCAGATGGTT	CTGACGAAGA	CAATAAATG	700
S S P	C T D	A L T C	I K A	D L F	C D G	V N C	P D G	S D E	N K M	220
TGTCCACAG	TTTGTGATGG	AAGATTGTTG	TAACTGGAT	CATCTGGGTC	TTTCCAGGCT	ACTCATTATC	CAAAACCTTC	TGAACAAGT	GTGTCTGCC	800
CAT	V C D G	R P L	L T G	S S G S	F O A	T H Y	P K P S	E T S	V V C	253
AGTGGATCAT	ACGTGTAAAC	CAAGGACTTT	CCATTAAACT	GAGCTTCGAT	GATTTTAATA	CATATTATAC	AGATATATTA	GATATTATG	AAGGTGTAGG	900
Q W I	R V N	O G L	S I K L	S F D	D F N	T Y Y T	D I Y	E N G	A G V G	287
ATCAAGCAAG	ATTTTAAGAG	CTTCTATTG	GGAACTAAT	CCTGGCACAA	TAAGAATTTT	TTCCAACCAA	GTTACTGCCA	CCTTCTTAT	AGAATCTGAT	1000
S S K	I L R	A S I W	B T N	P G T	I R I F	S N Q	V T A	T F L I	E S D	320
GAAAGTGATT	ATGTTGGCTT	TAATGCAACA	TATACCTGAT	TTAACAGCAG	TGAGCTTAAT	AATTATGAGA	AAATTAATTG	TAACTTTGAG	GATGGCTTTT	1100
E S D	Y V G F	N A T	Y T A	F N E S	E L N	N Y E	K I N C	N F E	D G F	353
GTTCCTGGGT	CCAGGACTCA	AATGATGATA	ATGAATGGGA	AAGGATTCAG	GGAAGCACTT	TTTCTCCTTT	TACTGGACCC	AATTTGACC	ACACTTTTGG	1200
C F W V	Q D L	N D D	N E W E	R I Q	G S T	F S P P	T G P	N F D	H T F P G	387
CAATGCTTCA	GGATTTTACA	TTTCTACCCC	AACTGGACCA	GGAGGGAGAC	AAGAAGCAGT	GGGGCTTTTA	AGCCTCCCTT	TGGACCCAC	TTTGAGGCCA	1300
N A S	G P Y	I S T P	T G P	G G R	Q E R V	G L L	S L P	L D P T	L E P	420
GCTTGCCTTA	GTTCCTGTA	TCATATGTAT	GGTGAAGATG	TCCAATAAT	AAGCATTAT	ATCAGCAATG	ACCAAAATAT	GGAGAAGACA	GTTTTCCAAA	1400
A C L	S F W Y	H M Y	G E N	V H K A	S I N	I S N	D O N M	E K T	V P Q	453
AGGAAGGAAA	TTATGGAGAC	AATGGAATT	ATGGACAAGT	AACCTTAAT	GAAACAGTTA	AATTTAAGGT	TGCTTTTAAT	GCTTTTAAAA	ACAAGATCCT	1500
K E G N	Y G D	N W N	Y G Q V	T L N	E T V	K F K V	A F N	A P K	N K I L	487
GAGTGATTT	GGTGTGGATG	ACATTAGCCT	AACATATGGG	ATTGTCAATG	GGAGTCTTTA	TCCAGAACCA	ACTTTGGTGC	CAACTCTCC	ACCAGAACTT	1600
S D I	A L D	D I S L	T Y G	I C N	G S L Y	P E P	T L V	P T P P	P E L	520
CCTACGACT	GTGGAGGACC	TTTGTAGCTG	TGGGAGCCAA	ATACAACATT	CAGTTTCTAG	AACTTTCCAA	ACAGCTACCC	TAACTGGGCT	TTCTGTGTTT	1700
P D F	C G G P	F E L	W E P	N T T F	S S T	N F P	N S Y P	N L A	F C V	553
GGATTTTAAA	TGCACAAAAA	GGAAAGAATA	TACAACCTCA	TTTCAAGAA	TTTGACTTAG	AAATATTAA	CGATGTAGTT	GAAATAAGAG	ATGGTGAAGA	1800
W I L N	A Q K	G K N	I Q L H	F Q E	F D L	S N I N	D V V	E I R	D G E E	587
AGCTGATTC	TTGCTCTTAG	CTGTGTACAC	AGGGCCTGGC	CCAGTAAGAAG	ATGTGTCTCT	TACCACCAAC	AGAATGACTG	TGCTTCTCAT	CACTAACGAT	1900
A D S	L L L	A V Y T	G P G	P V K	D V F S	T T N	R M T	V L L I	T N D	620
GTGTGGCAA	GAGGAGGTTT	TAAAGCAAC	TTTACTACTG	GCTATCACTT	GGGATTCCTA	GAGCCATGCA	AGCCAGACCA	TTTTTAAATG	AAAAATGGAG	2000
V L A	R G G F	K A N	F T T	G Y H L	G I P	E P C	K A D H	F Q C	K N G	653
AGTGTGTTC	ACTGTGTAAT	CTCTGTGAGG	GTCACTTGCA	CTGTGGAGAT	GGCTCAGATG	AAGCAGATTG	TGTGCGTTTT	TTCAATGGCA	CAACGAACAA	2100
E C V P	L V N	L C D	G H L H	C E D	G S D	B A D C	V R F	F N G	T T N N	687
CAATGGTTTA	GTGCGTTTCA	GAATCCAGAG	CATATGGCAT	ACAGCTTGCT	CTGGAAGCTG	GACCAACCCG	ATTTCAAATG	ATGTTTGTCA	ACTGCTGGGA	2200
N G L	V R F	I Q S	I W H	T A C	A E N W	T T Q	I S N	D V C Q	L L G	720
CTAGGGAGTG	GAAATCATC	AAAGCCAATC	TTCTCTACCG	ATGGTGGACC	ATTGTGCAAA	TTAAACACAG	CACCTGATGG	CCACTTAATA	CTAACACCCA	2300
L G S	G N S S	K P I	F S T	D G C P	F V K	L N T	A P D G	H L I	L T P	753
GTACACAGTG	TTTACAGTG	TCCTTGATTC	GRTTACAGTG	TAAACATAAA	TCTTTGTGAA	AAAACTGGC	AGCTCAAGAG	ATCACCCCAA	AGATTGTTGG	2400
S Q Q C	L Q D	S L L	R L Q C	N H K	S C G	K K L A	A Q D	I T P	K I V G	787
AGGAAGTAAT	GGCAAGAAG	GGCGCTGGCC	CTGGGTTGTT	GGTCTGTATT	ATGGCGGGCG	ACTGCTCTGC	GGCCACTCTC	TCGTCAGCAG	TGACTGGCTG	2500
G S N	A K E	G A W V	W V V	G L Y	Y G G R	L L C	G A S	L V S S	D W L	820
GTGTCGCGCG	CACACTGCGT	GTATGGGAGA	AACTTAGAGC	CATCCAAGTG	GACAGCAATC	CTAGGCGCTG	ATATGAAATC	AAATCTGACC	TCTCTCAAAA	2600
V S A	A H C V	Y G R	N L S	P S K W	T A I	L G L	H M K S	N L T	S P Q	853
CAGTCCCTCG	ATTAATAGAT	GAAATTGTCA	TAAACCTTCA	TTACAATAGG	CGAAGAAAGG	ACAAGGACAT	TGCCATGATG	CACTCTGGAAT	TAAAGTGAA	2700
T V P R	L I D	E I V	I N P H	Y N R	R R X	D N D I	A M H	H L E	F K V N	887
TTACACAGAT	TACATACAA	CTATTGTGTT	ACCGGAAGAA	AATCAAGTTT	TTCTTCCAGG	AAGAAATGTT	TCTATTCGCT	GTGCGGGGAC	GGTGTATAT	2800
Y T D	Y I Q	P I C L	P E E	N Q V	F P P G	R N C	S I A	G W G T	V V Y	920
CAAGGTAATA	CTGCAACAT	ATTGCAAGAA	GCTGATGTTT	CTCTTCTATC	AAATGAGAGA	TGCCAACACG	AGATGCCAGA	ATATAACATT	ACTGAAAATA	2900
Q G T	T A N I	L Q E	A D V	P L L S	N E R	C Q Q	O M P E	Y N I	T E N	953
TGATATGTGC	AGGCTATGAA	GAAGGAGGAA	TAGATTCTTG	TCAGGGGGAT	TCAGGAGGAC	CATTAAATGT	CCAAGAAAAC	AACAGGTGGT	TCCTTGCTGG	3000
N I C A	G Y E	E G G	I D S C	Q G D	S G G	L M C	Q E N	N R W	P L A G	987
TGTGACCTCA	TTTGATACA	AGTGTGCCCT	GCCTAATGCG	CCCGAGGTGT	ATGCCAGGGT	CTCAAGGTTT	ACCGAATGGA	TACAAAGTTT	TCTACATTAG	3100
V T S	F G Y	K C A L	P N R	P G V	Y A R V	S R F	T E W	I Q S F	L K H	1019
CGCATTTCTT	AACTAAACA	ATTATTTTCC	CATCTACTCT	TAGAAGCAT	GAAATTAAG	GGAAATTAAG	TGTTTCTGAC	AAAAATTTTA	AAAAGTTACC	3200
AAAGCTTTTT	ATTCTTACCT	ATGCTAATGA	AATGCTAGGG	GGCCAGGGAA	ACAAATTTT	AAAAATTAAT	AAATTCACCA	TAGCAATACA	GAATAACTTT	3300
AAATACCAT	TAAATACATT	TGATTTTCAT	TGTGAACAGG	TATTTCTTCA	CAGATCTCAT	TTTAAATTT	CTTAATGATT	ATTTTATTA	CTTACTGTG	3400
TTTAAAGGGA	TGTTATTTTA	AAGCATATAC	CATACACTTA	AGAAATTTTA	GCAGATTTTA	AAAAAGAAAG	AAAAATAATT	GTTTTCCCA	AAGTATGTCA	3500
CTGTGGAAA	TAAACTGCCA	TAAATTTTCT	AGTTCCAGTT	TAGTTTCTGT	CTATTAGCAG	AAACTCAATT	GTTCCTCTGT	CTTTCTATC	AAAAATTTCA	3600
ACATATGCAT	AACCTTAGTA	TTTTCCCAAC	CAATAGAAAC	TATTTATTGT	AAGCTTATGT	CACAGGCGCT	GACTAAATTT	ATTTTACGTT	CCTCTT	3696

FIGURE 2: Nucleotide and translated amino acid sequence of human enterokinase. Numbering at the right indicates the nucleotide or amino acid residue at the end of each line. Amino acids are shown in single-letter code. The termination codon is shown by an asterisk (*). The sequences contained in individual cDNA clones are as follows: HEK27, nt 1-2362; HEK19a, nt 948-2139; HEK18, nt 1451-2788; HEK12, nt 1591-3045; HEK6, nt 1762-2714; HEK3, nt 2278-2714; HEK1, nt 2454-3668; HEK5, nt 2511-3969.

Ho
Se
PoBo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
PoHo
Bo
Po

Hek	MGSKRgIeSR	MhSLeeYEIM	FaElFaILVv	LCAGLIAVSc	LtIkeSqrqA	AlQsSHEARa	TfKItSGVtY	NPhLQDKLSV	DPKVLAFDIO	QMIdelPISs	100
Bek	MGSKRvopSR	HrSLtEVEVM	FaVLvILVv	LCAGLIAVSV	LeIqSvYkda	AEGKSHEARg	TIKIISGaTy	NPhLQDKLSV	DPKVLAFDIO	QMIdIdFqSS	100
Pek	MGSKRIpSR	HrSLeIyEVM	FtAlFaILmV	LCAGLIAVSV	LtIkgSekda	AlGKSHEARg	TaKItSGVtY	NPhLQDKLSV	DPKVLAFDIO	QMIdelFqSS	100
.....											
Hek	NLKNEYKNSR	VLOPENGII	VvFDLIFaQW	VSDQNVKEEL	IQGIEANKSS	QLVtPHIDIN	SVdII.....	dKLTtTeshla	TPGNVIEECI	185	
Bek	NLKNEYKNSR	VLOPENGII	VvFDLIFaQW	VSDQNVKEEL	IQGIEANKSS	QLVtPHIDIN	SIDItaslea	fatispette	aKLTtTeshla	TPGNVIEECI	200
Pek	NLKNEYKNSR	VLOPENGVI	VvFDLIFaQW	VSDNIKEEL	IQGIEANKSS	QLVtPHIDIN	SIDItaslea	yattspette	dKLTtTeshla	TPGNVIEECI	200
.....											
Hek	PgSspCTDAL	tCIkaDLFCD	GEVNCPOGSD	EDnKNCATvC	DGrFLLTgSS	GSFqathYPR	pS.etSVVCq	WIIRVNOGLS	IklEpddPNT	YytDIdLIYE	284
Bek	PdSrlCaDAL	kCIaIdLFCD	GEVNCPOGSD	EDnKNCATvC	DGrFLLTgSS	GSFqathYPR	pS.SaLlRaVCT	WIIRVNOGLS	IqlnFdyPNT	YyaDvLnIYE	300
Pek	PgSrpCaDAL	kCIaIdLFCD	GEVNCPOGSD	EDnKNCATvC	DGrFLLTgSS	GSFqathYPR	pS.SaLlRaVCT	WIIRVNOGLS	IqlnFdyPNT	YyaDvLnIYE	299
.....											
Hek	GvGSSKILRA	SIWetNPGCI	RIFSNOVTaT	FLIEsDEsDY	VGFpaTYTAP	NSaELNMYEK	INCNPEDGFC	FWvQDLNDON	EWERIOGstF	SPtTGpFDH	384
Bek	GvGSSKILRA	SIWetNPGCI	RIFSNOVTaT	FLIEsDEsDY	VGFpaTYTAP	NSaELNMYEK	INCNPEDGFC	FWvQDLNDON	EWERIOGstF	SPtTGpFDH	400
Pek	GvGSSKILRA	SIWetNPGCI	RIFSNOVTaT	FLIEsDEsDY	VGFpaTYTAP	NSaELNMYEK	INCNPEDGFC	FWvQDLNDON	EWERIOGstF	SPtTGpFDH	399
.....											
Hek	TFCNAGFYI	STPTGPGGRQ	ERVGLLsLPL	dPTIEpaCLS	FWYhMYGENV	hKLSINISnd	ONaEktVFOK	EGNYGdNMNY	QOVTLNtEtk	FKVaFnaPKN	484
Bek	TFCNAGFYI	STPTGPGGRQ	ERVGLLsLPL	dPTIEpaCLS	FWYhMYGENV	hKLSINISnd	ONaEktVFOK	EGNYGdNMNY	QOVTLNtEtk	FKVaFnaPKN	500
Pek	TFCNAGFYI	STPTGPGGRQ	ERVGLLsLPL	dPTIEpaCLS	FWYhMYGENV	hKLSINISnd	ONaEktVFOK	EGNYGdNMNY	QOVTLNtEtk	FKVaFnaPKN	499
.....											
Hek	kILSDIALDD	ISLTyGICN	gLYPEPTLVP	TpPELPTDC	GGPILWEFN	TFPstNFPN	gYPNlAPCVW	ILNAQKGKNI	QLHFqEFdLE	NINDVVEIRD	584
Bek	kILSDIALDD	ISLTyGICN	gLYPEPTLVP	TpPELPTDC	GGPILWEFN	TFPstNFPN	gYPNlAPCVW	ILNAQKGKNI	QLHFqEFdLE	NINDVVEIRD	600
Pek	kILSDIALDD	ISLTyGICN	gLYPEPTLVP	TpPELPTDC	GGPILWEFN	TFPstNFPN	gYPNlAPCVW	ILNAQKGKNI	QLHFqEFdLE	NINDVVEIRD	599
.....											
Hek	GEeDSLILA	VYTGPpGVd	VFTTNRMtV	LlITndvLar	gGFKANETTG	YhLGIPePK	aDhFOCKnGE	CvPLVNLCDG	hIHCeGDSDE	AhcVrIINGT	684
Bek	GEeDSLILA	VYTGPpGVd	VFTTNRMtV	LlITndvLar	gGFKANETTG	YhLGIPePK	aDhFOCKnGE	CvPLVNLCDG	hIHCeGDSDE	AhcVrIINGT	700
Pek	GEeDSLILA	VYTGPpGVd	VFTTNRMtV	LlITndvLar	gGFKANETTG	YhLGIPePK	aDhFOCKnGE	CvPLVNLCDG	hIHCeGDSDE	AhcVrIINGT	699
.....											
Hek	tnnnGLVrPR	IOSIWHtACA	SNYTOISnd	VQQLGLGSG	HSBkPIFstc	GGPVLKNTA	PdGhLILtPS	QCLQLDSLIR	LCQNHKSCGK	KlaaOdIdPK	784
Bek	tdssGLVrPR	IOSIWHtACA	SNYTOISnd	VQQLGLGSG	HSBkPIFstc	GGPVLKNTA	PdGhLILtPS	QCLQLDSLIR	LCQNHKSCGK	KlaaOdIdPK	800
Pek	tnnnGLVrPR	IOSIWHtACA	SNYTOISnd	VQQLGLGSG	HSBkPIFstc	GGPVLKNTA	PdGhLILtPS	QCLQLDSLIR	LCQNHKSCGK	KlaaOdIdPK	799
.....											
Hek	IVGGenakeG	AWPwVvLy	ggrllCGASL	VsEdWLVSAA	HCvYGRNIEP	SKWtAILGLH	MKSILTSPOI	vPLRIdEIVI	NPHYNRRRKd	nDIAMHLEf	884
Bek	IVGGSdsrEG	AWPwVvLy	ggrllCGASL	VsEdWLVSAA	HCvYGRNIEP	SKWtAILGLH	MKSILTSPOI	vPLRIdEIVI	NPHYNRRRKd	nDIAMHLEf	900
Pek	IVGGSdsrEG	AWPwVvLy	ggrllCGASL	VsEdWLVSAA	HCvYGRNIEP	SKWtAILGLH	MKSILTSPOI	vPLRIdEIVI	NPHYNRRRKd	nDIAMHLEf	899
.....											
Hek	KVNIXDYIOP	ICLPEENQVF	PPGRICSIAG	WGtVvYOGtC	AnILOEADVP	LlSNERCQOO	MPEYNITENM	ICAGYEAEGGI	DSOQGDSSGP	LNCqENNRWf	984
Bek	KVNIXDYIOP	ICLPEENQVF	PPGRICSIAG	WGtVvYOGtC	AnILOEADVP	LlSNERCQOO	MPEYNITENM	ICAGYEAEGGI	DSOQGDSSGP	LNCqENNRWf	1000
Pek	KVNIXDYIOP	ICLPEENQVF	PPGRICSIAG	WGtVvYOGtC	AnILOEADVP	LlSNERCQOO	MPEYNITENM	ICAGYEAEGGI	DSOQGDSSGP	LNCqENNRWf	999
.....											
Hek	LAGVTSFCYk	CALPNRPGVY	ARVvPTeWI	QSFLH	1019						
Bek	LAGVTSFCYk	CALPNRPGVY	ARVvPTeWI	QSFLH	1035						
Pek	LAGVTSFCYk	CALPNRPGVY	ARVvPTeWI	QSFLH	1034						

FIGURE 3: Alignment of human (Hek), bovine (Bek) (Kitamoto et al., 1994), and porcine (Pek) (Matsushima et al., 1994) enterokinase amino acid sequences. Amino acids are shown in single-letter code. Residues that are identical in all three species are in capital letters; unconserved residues are in lower case. Numbering at the right refers to the translated amino acid sequence of each species of enterokinase. Cysteine residues are in boldface type. Potential N-linked glycosylation sites are in boldface underlined type. The potential signal anchor sequence is double underlined. The location of a potential alternatively spliced exon in bovine enterokinase is indicated by a dotted underline. This segment is notably variable among the aligned species. Sequence motifs in the heavy chain are indicated by numbered underlines that correspond to the domains shown in Figure 1.

over 599 nt. A similar degree of sequence identity is apparent when either the human or bovine enterokinase sequences are compared to the porcine enterokinase cDNA sequence (Matsushima et al., 1994).

Structural Features of Human Enterokinase. Most structural elements of human enterokinase are highly conserved (Figure 3). The similarities among the human, bovine, and porcine enterokinase sequences suggest that the mature proteins consist of two polypeptide chains derived by processing of a single-chain precursor. A potential myristoylation site is present at Gly2 (Rudnick et al., 1993). Amino acid residues 19–43 are hydrophobic and may constitute a signal-anchor sequence. The putative heavy chain contains six sequence motifs that appear to be homologous to four types of domains found in other proteins (Figure 4). As reported previously (Kitamoto et al., 1994), the cleavage site after Lys784 separates the heavy and light chains of enterokinase, and the light chain is homologous to the trypsin-like family of serine proteases. In all three cloned enterokinases, the sequence surrounding this cleavage site is consistent with the known substrate specificity of trypsin.

Enterokinase domains 1 and 5 are homologous to cysteine-rich repeats in the low-density lipoprotein receptor (Sudhof

et al., 1985); domain 6 is homologous to a segment of the macrophage scavenger receptor (Freeman et al., 1990), as reported previously (Kitamoto et al., 1994).

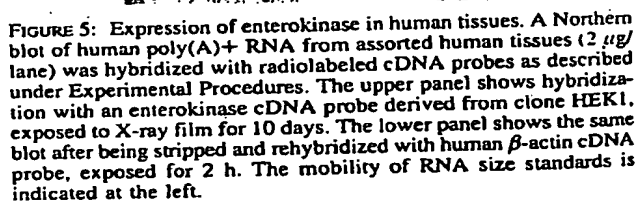
During the analysis of the bovine enterokinase sequence (Kitamoto et al., 1994) domain 4 was recognized as a member of a sequence family that includes two motifs identified first in complement component C1r (Leytus et al., 1986). Domain 2 of porcine enterokinase then was found to belong to the same sequence family (Matsushima et al., 1994). As indicated in Figures 3 and 4, two C1r/s domains clearly are present in human, bovine, and porcine enterokinase.

Domain 3 of bovine enterokinase (Kitamoto et al., 1994) was recognized as homologous to segments of the metalloproteases meprin A (Jiang et al., 1992) and meprin B (Johnson & Hersh, 1992) and to a domain of the A5 protein of *Xenopus laevis* (Takagi et al., 1991). The name "meprin domain" was suggested for this motif (Kitamoto et al., 1994). However, a previous report had described the same motif in meprins, the *Xenopus* A5 protein, and in the extracellular domain of receptor protein tyrosine phosphatase μ (Gebbinck et al., 1991); the name "MAM" domain was proposed (for "meprin", "A5", and "mu") (Beckmann & Bork, 1993). The

FIGURE 4: Alignment of C1r/s domains of human enterokinase. Human enterokinase domains Hek-2 and Hek-4 are numbered as in Figures 1 and 3. Domains Hek-2 and Hek-4 are aligned with selected ≈ 120 amino acid repeats from the *Drosophila melanogaster* tollid protein (Shimell et al., 1991) and from complement component C1r (Leytys et al., 1986). The significance of gap alignments was evaluated by comparing the optimal alignment score to the mean and SD of scores obtained for 30 alignments of randomized sequences, using the normal distribution to estimate the probability (P) that the alignment could occur by chance. The value obtained for P was $<10^{-14}$.

The function of the enterokinase heavy-chain domains is unknown. Related domains in other proteins appear to bind ligands or mediate protein-protein interactions. For example, the α -subunit of mouse meprin A associates with the β -subunit, possibly through MAM domains in each subunit. This association is required for membrane localization of the mature α -subunit, which lacks a membrane-spanning domain (Marchant et al., 1994). Thus, the MAM domain of enterokinase could interact with other proteins that contribute to membrane localization or enzyme activity. A role for the heavy chain in determining substrate specificity would be consistent with the reported ability of heating (Barns & Elmslie, 1974; Anderson et al., 1977), acetylation of amino groups (Baratti & Maroux, 1976), or dissociation of the light chain by partial reduction (Light & Fonseca, 1984) to selectively impair enterokinase activity toward trypsinogen without markedly affecting activity toward small amides or esters.

Human and porcine enterokinase also lack one amino acid residue that is found in bovine enterokinase domain 2 (Figure 3); this deletion removes two possible concatenated N-linked glycosylation sites. Several additional glycosylation sites are not conserved, so that human, bovine, and porcine enterokinase heavy chains have 14, 17, and 18 potential N-glycosylation sites, respectively.



Chromosome Localization of the Human Enterokinase Gene. Fluorescent *in situ* hybridization was used to physically localize the human enterokinase gene. To obtain an adequate hybridization signal, the inserts of cDNA clones HEK1 and HEK6 were mixed, thereby including ≈ 1.9 kb of the cDNA sequence. The DNA was labeled with biotin

FIGURE
shows
shows

and h
some
cyana
cein i
Fifty
repre
hybrid
on ch
was d
staini
hybrid
The
gene
al., 1

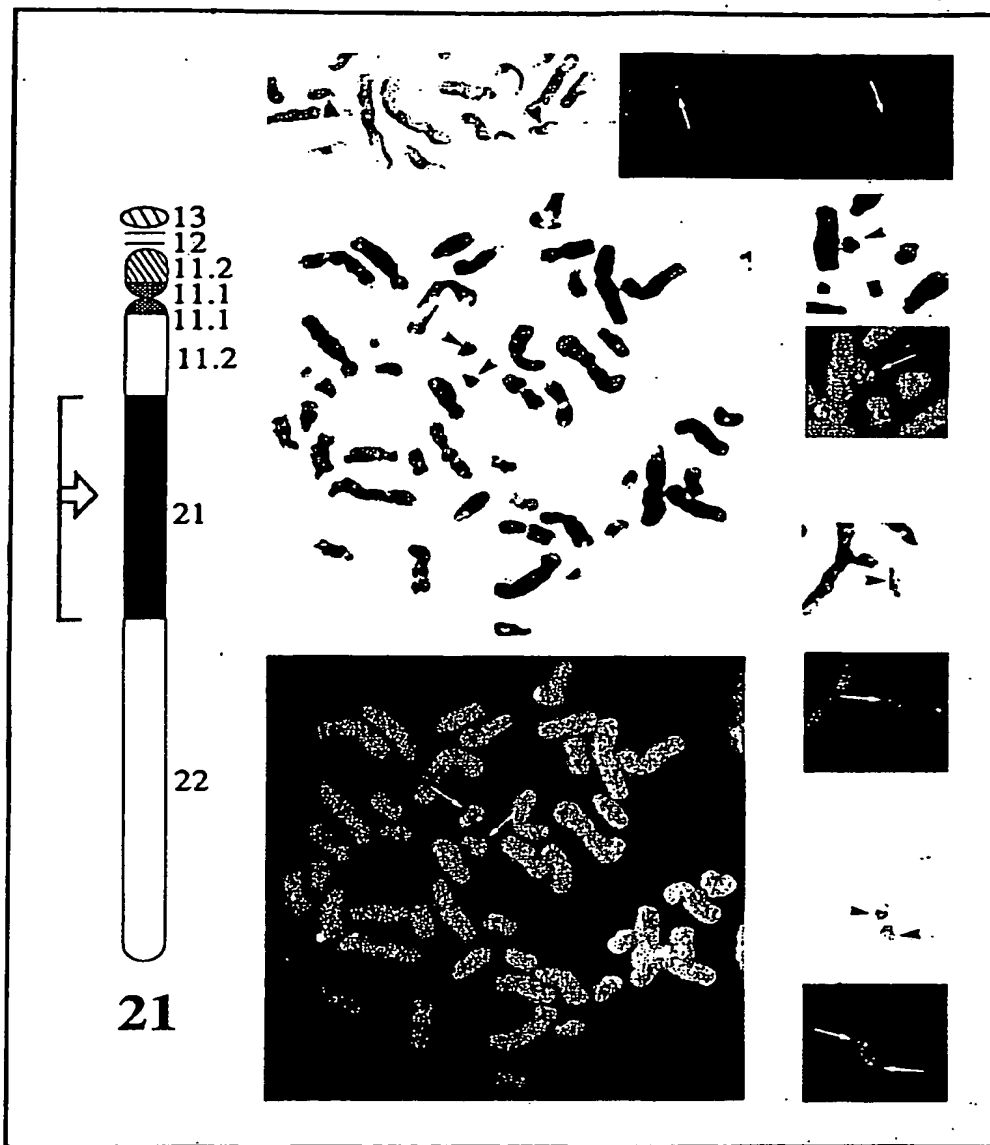


FIGURE 6: Fluorescent *in situ* hybridization localization of the enterokinase gene to human chromosome 21q21. Five metaphase spreads are shown. Arrows indicate biotin-labeled probe hybridization (color) and the position of the same spreads banded using Giemsa dye. Also shown is an idiogram of chromosome 21 with band q21, to which the probes hybridize, indicated by an arrowhead.

and hybridized to prometaphase spreads of human chromosomes. Labeled DNA was detected with fluorescein isothiocyanate-conjugated avidin DCS and amplified with fluorescein isothiocyanate-conjugated goat anti-avidin D antibodies. Fifty independent metaphase spreads were analyzed, and five representative spreads are shown (Figure 6). Specific hybridization of the enterokinase cDNA probe was observed on chromosome 21; no consistent secondary hybridization was detected. 4,6-Diamidino-2-phenylindole dihydrochloride staining and Giemsa banding confirmed the location of the hybridization signals on chromosome 21 band q21.

The human enterokinase locus appears to be close to the gene for β -amyloid precursor protein at 21q21.2 (Nizetic et al., 1994), which is mutated in one form of inherited

Alzheimer disease (Goate et al., 1991), and to the gene for superoxide dismutase at 21q22.1, which is mutated in familial amyotrophic lateral sclerosis (Rosen et al., 1993). Enterokinase also is in or near a region implicated in specific features of Down syndrome, although the precise locations of chromosome 21 segments that contribute to Down syndrome remain unknown (Korenberg et al., 1994). The cloning of cDNA for human enterokinase will enable fine structure physical and genetic mapping of these loci and the characterization of mutations in congenital enterokinase deficiency (Hadorn et al., 1969; Haworth et al., 1971). These clones also facilitate the study of biosynthetic targeting to apical brush border membranes, zymogen activation, and substrate specificity of human enterokinase.

ACKNOWLEDGMENT

We thank Lisa Westfield for synthesis of oligonucleotides and Drs. Xin Yuan and Qingyu Wu for many helpful discussions.

REFERENCES

- Anderson, L. E., Walsh, K. A., & Neurath, H. (1977) *Biochemistry* 16, 3354-3360.
- Baratti, J., & Maroux, S. (1976) *Biochim. Biophys. Acta* 452, 488-496.
- Baratti, J., Maroux, S., Louvard, D., & Desnuelle, P. (1973) *Biochim. Biophys. Acta* 315, 147-161.
- Barns, R. J., & Elmslie, R. G. (1974) *Biochim. Biophys. Acta* 350, 495-498.
- Beckmann, G., & Bork, P. (1993) *Trends Biochem. Sci.* 18, 40-41.
- Dangott, L. J., Jordan, J. E., Bellet, R. A., & Garbers, D. L. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 2128-2132.
- Davie, E. W., & Neurath, H. (1955) *J. Biol. Chem.* 212, 515-529.
- Feinberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem.* 132, 6-13.
- Fonseca, P., & Light, P. (1983) *J. Biol. Chem.* 258, 14516-14520.
- Freeman, M., Ashenas, J., Rees, D. J. G., Kingsley, D. M., Copeland, N. G., Jenkins, N. A., & Krieger, M. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 8810-8814.
- Gebbink, M. F. B. G., van Elten, I., Hateboer, G., Suijkerbuijk, R., Beijersbergen, R. L., Geurts van Kessel, A., & Moolenaar, W. H. (1991) *FEBS Lett.* 290, 123-130.
- Gish, W., & States, D. J. (1993) *Nature Genet.* 3, 266-272.
- Goate, A., Chartier-Harlin, M.-C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N., James, L., Mant, R., Newton, P., Rooke, K., Roques, P., Talbot, C., Pericak-Vance, M., Roses, A., Williamson, R., Rossor, M., Owen, M., & Hardy, J. (1991) *Nature* 349, 704-706.
- Gunning, P., Ponte, P., Okayama, H., Engel, J., Blau, H., & Kedes, L. (1983) *Mol. Cell. Biol.* 3, 787-795.
- Hadorn, B., Tarlow, M. J., Lloyd, J. K., & Wolff, O. H. (1969) *Lancet* i, 812-813.
- Haworth, J. C., Gourley, B., Hadorn, B., & Sumida, C. (1971) *J. Pediatr.* 78, 481-490.
- Jiang, W., Gorbea, C. M., Flannery, A. V., Beynon, R. J., Grant, G. A., & Bond, J. S. (1992) *J. Biol. Chem.* 267, 9185-9193.
- Jiang, Y.-P., Wang, H., D'Eustachio, P., Musacchio, J. M., Schlessinger, J., & Sap, J. (1993) *Mol. Cell. Biol.* 13, 2942-2951.
- Johnson, G. D., & Hersch, L. B. (1992) *J. Biol. Chem.* 267, 13505-13512.
- Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W., & Sadler, J. E. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 7588-7592.
- Korenberg, J. R., Chen, X.-N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, C., Dignan, P., Distèche, C., Graham, J. M., Jr., Hudgins, L., McGillivray, B., Miyazaki, K., Ogasawara, N., Park, J. P., Pagon, R., Pueschel, S., Sack, G., Say, B., Schuffenhauer, S., Soukup, S., & Yamanaka, T. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 4997-5001.
- Kunitz, M. (1939) *J. Gen. Physiol.* 22, 429-446.
- LaVallie, E. R., Rehemtulla, A., Racie, L. A., DiBlasio, E. A., Ferenz, C., Grant, K. L., Light, A., & McCoy, J. M. (1993) *J. Biol. Chem.* 268, 23311-23317.
- Leytus, S. P., Kurachi, K., Sakariassen, K. S., & Davie, E. W. (1986) *Biochemistry* 25, 4855-4863.
- Lichter, P., Cremer, T., Borden, J., Manuelidis, L., & Ward, D. C. (1988) *Hum. Genet.* 80, 224-234.
- Liepnies, J. J., & Light, A. (1979) *J. Biol. Chem.* 254, 1677-1683.
- Light, A., & Fonseca, P. (1984) *J. Biol. Chem.* 259, 13195-13198.
- Lojda, Z., & Malis, F. (1972) *Histochemie* 32, 23-29.
- Magee, A. I., Grant, D. A. W., & Hermon-Taylor, J. (1981) *Clin. Chim. Acta* 115, 241-254.
- Marchant, P., Tang, J., & Bond, J. S. (1994) *J. Biol. Chem.* 269, 15388-15393.
- Matsumura, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., Umeyama, H., Inoue, H., Takahashi, T., & Takahashi, K. (1994) *J. Biol. Chem.* 269, 19976-19982.
- Miyoshi, Y., Onishi, T., Sano, T., & Komii, N. (1990) *Gastroenterol. Jpn.* 25, 320-327.
- Naude, R. J., Da Silva, D., Edge, W., & Oelofsen, W. (1993) *Comp. Biochem. Physiol.* 105B, 591-595.
- Nizetic, D., Gellen, L., Hamvas, R. M. J., Mott, R., Grigoriev, A., Vatcheva, R., Zehetner, G., Yaspo, M.-L., Dutriaux, A., Lopes, C., Delabar, J.-M., Van Broeckhoven, C., Potier, M.-C., & Lehrach, H. (1994) *Hum. Mol. Genet.* 3, 759-770.
- Pavlov, I. P. (1902) *The Work of the Digestive Glands*, 1st ed., Charles Griffin & Co., London.
- Rigby, P. W. J., Dieckmann, M., Rhodes, C., & Berg, P. (1977) *J. Mol. Biol.* 113, 237-251.
- Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J. P., Deng, H.-X., Rahmani, Z., Krizus, A., McKenna-Yasek, D., Cayabyab, A., Gaston, S. M., Berger, R., Tanzi, R. E., Halperin, J. J., Herzfeldt, B., Van den Bergh, R., Hung, W.-Y., Bird, T., Deng, G., Mulder, D. W., Smyth, C., Laing, N. G., Soriano, E., Pericak-Vance, M. A., Haines, J., Rouleau, G. A., Gusella, J. S., Horvitz, H. R., & Brown, R. H., Jr. (1993) *Nature* 362, 59-62.
- Rudnick, D. A., McWherter, C. A., Gokel, G. W., & Gordon, J. I. (1993) *Adv. Enzymol.* 67, 375-430.
- Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989), in *Molecular Cloning: A Laboratory Manual*, 2nd ed., pp 387-389, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463-5467.
- Shimell, M. J., Ferguson, E. L., Childs, S. R., & O'Connor, M. B. (1991) *Cell* 67, 469-481.
- Sudhof, T. C., Goldstein, J. L., Brown, M. S., & Russell, D. W. (1985) *Science* 228, 815-822.
- Takagi, S., Hirata, T., Agata, K., Mochii, M., Eguchi, G., & Fujisawa, H. (1991) *Neuron* 7, 295-307.
- Yamashina, I. (1956) *Acta Chem. Scand.* 10, 739-743.
- Yunis, J. J. (1976) *Science* 191, 1268-1270.

BI9500111

Exhibit 19

A Novel Trypsin-like Serine Protease (Hepsin) with a Putative Transmembrane Domain Expressed by Human Liver and Hepatoma Cells†

Steven P. Leytus,¹ Keith R. Loeb,^{1,2} Frederick S. Hagen,¹ Kotoku Kurachi,^{1,2} and Earl W. Davie*¹

Department of Biochemistry, University of Washington, Seattle, Washington 98195, and ZymoGenetics, Inc., 2121 North 35th Street, Seattle, Washington 98103

Received August 24, 1987; Revised Manuscript Received October 16, 1987

ABSTRACT: Recombinant clones with cDNA inserts coding for a new serine protease (hepsin) have been isolated from cDNA libraries prepared from human liver and hepatoma cell line mRNA. The total length of the cDNA is approximately 1.8 kilobases and includes a 5' untranslated region, 1251 nucleotides coding for a protein of 417 amino acids, a 3' untranslated region, and a poly(A) tail. The amino acid sequence coded by the cDNA for hepsin shows a high degree of identity to pancreatic trypsin and other serine proteases present in plasma. It also exhibits features characteristic of zymogens to serine proteases in that it contains a cleavage site for protease activation and the highly conserved regions surrounding the His, Asp, and Ser residues that participate in enzyme catalysis. In addition, hepsin lacks a typical amino-terminal signal peptide. Hydrophathy analysis of the protein sequence, however, revealed a very hydrophobic region of 27 amino acids starting 18 residues downstream from the apparent initiator Met. This region may serve as an internal signal sequence and a transmembrane domain. This putative transmembrane domain could be involved in anchoring hepsin to the cell membrane and orienting it in such a manner that its carboxyl terminus, containing the catalytic domain, is extracellular.

Many biological processes which require specific, limited proteolysis are mediated by a member(s) of the serine protease family of proteolytic enzymes. These proteases exist as single- or two-chain zymogens that are activated by specific and limited proteolytic cleavage (Neurath & Walsh, 1976). They contain three principal active-site amino acids (His, Asp, and Ser) that participate in peptide bond hydrolysis (Blow et al., 1969). In addition, they share considerable structural similarities in their catalytic chains.

Among the best-studied serine proteases are those that are found in plasma. These enzymes are involved in processes such as blood coagulation (Davie et al., 1979), fibrinolysis (Christman et al., 1977; Collen, 1980), and complement activation (Reid & Porter, 1981). The active form of most plasma serine proteases consists of two polypeptide chains held together by a disulfide bond(s), a highly conserved catalytic chain derived from the carboxyl-terminal end of the precursor polypeptide, and a unique noncatalytic chain derived from the amino-terminal portion of the polypeptide chain. The presence of a noncatalytic chain(s) distinguishes the plasma serine proteases from the digestive proteases of the pancreas. By mediating interactions with other proteins or surfaces, non-catalytic chains influence the action of plasma serine proteases on their selected substrates. The biosynthesis of most of the serine proteases present in plasma occurs in the liver. Although at least 20 different serine proteases synthesized in the liver have been described thus far, it is quite likely that many more exist.

Recent reports have identified a number of new serine proteases produced in different tissues and cell types. Cook

et al. (1985, 1987) have described a cDNA coding for a new serine protease that is expressed during adipocyte differentiation. Gershenfeld and Weissman (1986) and Lobe et al. (1986) have cloned cDNAs coding for new serine proteases expressed by cytotoxic T lymphocytes. Newly characterized proteins have also been isolated from cytotoxic T lymphocytes (Pasternack et al., 1986; Young et al., 1986; Masson & Tschopp, 1987), liver (Tanaka et al., 1986), ovary (Eisenhauer & McDonald, 1986), pituitary gland (Cromlish et al., 1986), embryo fibroblast cells (Billings et al., 1987), seminal plasma (Watt et al., 1986), submaxillary gland (Lundgren et al., 1984), and tumor cells (LaBombardi et al., 1983) that exhibit properties typical of serine proteases. Additional new proteases have been reported, but not all have been identified as belonging to the serine protease family. Although the majority of serine proteases are synthesized with signal peptides that direct their secretion outside of the cell, some of the new serine proteases recently reported may be associated with cell membranes (LaBombardi et al., 1983; Tanaka et al., 1986).

As a general approach to isolating cDNAs coding for serine proteases synthesized in the liver, a strategy was chosen that involved screening a human liver cDNA library with a synthetic oligodeoxynucleotide probe coding for a highly conserved amino acid sequence known to exist in a number of different serine proteases. In this manner, recombinant clones were isolated that contained cDNA inserts coding for serine proteases synthesized in the liver, including human factor IX (Kurachi & Davie, 1982), prothrombin (Degen et al., 1983), and complement C1r (Leytus et al., 1986a). In this paper, we report the isolation and characterization of the cDNA coding for a new trypsin-like serine protease. This hepatocyte-expressed protease has been called hepsin.

EXPERIMENTAL PROCEDURES

DNA restriction endonucleases and DNA modification enzymes were purchased from Bethesda Research Laboratories or New England Biolabs. ³²P-Labeled nucleotides used in nick-translating cDNA fragments (Maniatis et al., 1982) and 5'-end-labeling synthetic oligodeoxynucleotides (Maxam &

[†] This work was supported in part by research grants (HL 16919 and HL 31511) and a postdoctoral fellowship (GM 09118 to S.P.L.) from the National Institutes of Health.

¹ University of Washington.

² Present address: Department of Biochemistry, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226.

³ ZymoGenetics, Inc.

⁴ Present address: Department of Human Genetics, 4708 Medical Science II, University of Michigan Medical School, Ann Arbor, MI 48109.

CGCTTTCACAGGGAACCTACTGAGGGGACAGAGGTGAGGCAGCCTGGCCTAGCAGGGCCACAGCAGCAGCTCTGCTCAGGCGGCGCGCTCTGCGGGGCCACCATGCTCTGCCCA

127 GGCTGGAGACTGACCCGACCCCGGCACTACTGAGGCTCCGCGCCACCTGCTGGACCCAGCTTCCACACCTGGCCAGGAGGTGACGACAGGGAATCATTAACAAGAGGCACTGAC

1 246 M A Q K E G G R T V P C C S R P K V A A L T A G T L L L L T
ATG GCG CAG AAG GAG GGT GGC DGG ACT GTG CCA TGC TGC TCC AGA CCG AAG GTG GCA GCT CTC ACT GCG GCG ACC CTG CTA CTT CTG ACA

31 A I C G G A S W A T V A V L L L R S D Q E G P L Y P V Q V S S A D
338 GCG ATG GCG GCG GCA TCG TGG GCC ATT GTG GCT GTT CTC CTC AGG AGT GAC CAG GAG CCG CTG TAC CCA GTG CAG GTC AGC TCT GCG GAC

61 A R L M V F D K T E G T W R L L C S S R S N A R V A G L S C
426 GCT CGG CTC ATG GTC TTT GAC AAG ACG GAA GGG ACG TGG CCG CTG CTG TGC TCC TCG CCG TCC AAC GCC AGG GTA GCC GGA CTC AGC TGC

91 E E M G F L R A L T H S E L D V T R T G G G C C A N G G C T S G F F C C
518 GAG GAG ATG GCG TTC CTC AGG GCA CTG ACC CAC TCC GAG CTG GAC GTG CGA ACG CCG GCC AAT GGC ACG TCG GCG TTC TTC TGT GTG

121 D E G G R L F H T Q C R L L E V I S V C D C P R G R F L A A I C
606 GAC GAG GCG AGG CTG CCC CAC ACC CAG AGG CTG CTG GAG GTC ATC TCC GTG TGT GAT TGC CCC AGA GCC CGT TTC TTG GCC GCC ATC TGC

151 Q D C G R R K L P V D R I V G G C R D T C S L G C R W P H Q V S T
698 CAA GAC TGT GCG CGC AGG AAG CTG CCC GTG GAC CGC ATC GTG GGA GCG CGG CAC ACC AGC TTG GCC CGG TGG CCG TGG CAA GTG AGC CTT

181 R Y D G A H L C G G S L L S G G D W V L T A C C C C F P P Z R N R G
786 CGC TAT GAT GGA GCA CAC CTC TGT GGG GGA TCC CTG CTC TCC GGG GAC TGG GTG CTG ACA GCC CAC TCC TTC CCG GAG CCG AAC CGG

211 V L S R W R V F A G A V A Q A S P B G L Q L G T Q A V V Y Y H
876 GTC CTG TCC CGA TGG CGA GTG TTT GCG GGT GCC GTG GCC CAG GCC TCT CCC CAC GGT CTG CAG CTG GGG GTG CAG GCT GTG GTC TAC CAC

241 G G C Y L P P R D P N S E E N S B D I A C L V H L S S F L P L T
968 GGG GCG TAT CTT CCC TTT CGG GAC OCC AAC AGC GAG AAC AGC AAC GAT ATT GGC ATG GTC CAC CTC TCC AGT CCC CTG CCC CTC ACA

271 E Y I Q P V C L P A A G Q A L V D G K I C T V T G W G N T Q
1056 GAA TAC ATC CAG CCT GTG TGC CTC CCA GCT GCC GGC CAG GCC CTG GTG GAT GGC AAG ATC TGT ACC GTG ACG GGC TGG GGC AAC ACG CAG

301 Y Y T G C Q Q A G G V L Q E A R G V P I I T S N D G T C N G A D T F Y G A N
1146 TAC TAT GCG CAA CAG GCC GGG GTA CTC CAG GAG GCT CCA CGA GTG CCC ATA ATC AGC AAT GAT GTC TGC AAT GGC GCT GAC TTC YAT GGA AAC

331 Q I K P K M F C A G Y P E G G C I D A C Q G D S G G P P V C E
1236 CAG ATC AAG CCG AAG ATG TTC TGT GCT GGC TAC CCC GAG GGT GCG ATT GAC GCC TGC CAG GCC GAC AGC GGT GGT CCC TTT GTG TGT GAG

361 D S I S R T P R W R L C G I V S W G T G C T G C A L A Q A K P G V Y
1326 GAC AGC ATC TCT CCG ACG CCA CGT TGG CCG CTG TGT GGC ATT GTG AGT TGG GGC ACT TGC TGT GCC CCA CTG GCC CAG AAG CCA GGC GTC TAC

391 T K V S D F R E W I P Q A I K T H S E A S B H V T Q L *
1416 ACC AAA GTC AGT GAC TTC CCG GAG TGG ATC TTC CAG GCC ATA AAG ACT CAC TCC GAA GCC AGC GGC ATG GTG ACC CAG CTC TGA CCGGTGG

1507 CTCTCTGCTGGCAGCCTCAGGGGCCGAGGTGATCCCGGTGGTGGGATCCACGCTGGGCGGAGGATGGGACGTTTTCCTCTCTGGGCGCGGTCACAGGTCCAAGGACACCCCTCCCTC

1626 CAGGGTCTCTCTTCCACAGTGGCGGGCCCACTCAGCCCCGAGACCACCCAACTCAOCCCTCTGACCCCATGTAAATATTGTCTCTGCTGTCTGGGACTCTGTCTYAGGTGCCCTTGA

1745 TGATGGGATGCTCTTTAAATAATAAAGATGGTTTTGATT-poly(A)

FIGURE 2: Nucleotide sequence of the cDNA coding for human hepsin. The sequence was determined by analysis of the cDNA inserts shown in Figure 1. The predicted amino acid sequence is shown above the DNA sequence. The solid, inverted triangle marks the location of the inserted sequence found in clones HepG2UW17 and HepG2UW2 (see Figure 1). This sequence is not included in Figure 2. The boxed amino acid sequence represents a potential transmembrane domain. The solid arrow identifies an Arg-Ile bond that is probably cleaved when the inactive zymogen is converted to an active protease. The active-site His, Asp, and Ser residues are circled. The underlined nucleotide sequence is the site responsible for hybridizing to the synthetic oligodeoxynucleotide probe.

Digestion of the recombinant phage DNAs with *EcoRI* released inserts that ranged in size from approximately 800 to 1800 base pairs (bp). Two of these inserts (HepG2UW7 and HepG2UW20) were selected for further analysis. A 160 bp *EcoRI*-*XhoI* fragment derived from the extreme 5' end of HepG2UW7 was then employed as a hybridization probe, and the original 70 positives were rescreened. Subsequently, five additional clones, designated HepG2UW2, HepG2UW17, HepG2UW19, HepG2UW61, and HepG2UW63, were also selected for DNA sequence analysis. A restriction enzyme map for the seven cDNA inserts obtained from the Hep G2 library is shown in Figure 1. The strategy used to determine the cDNA sequence of hepsin from the various clones is also described in Figure 1.

The complete nucleotide sequence of the cDNA coding for hepsin is shown in Figure 2, along with the deduced amino

acid sequence. The total length of the cDNA was 1783 bp. This is consistent with the size of the mRNA for hepsin present in Hep G2 cells as determined by Northern blot analysis (data not shown). The cDNA includes 245 nucleotides of untranslated sequence at the 5' end, 1251 nucleotides coding for a protein of 417 amino acids, a stop codon of TGA, and 284 nucleotides of untranslated sequence at the 3' end. The ATG codon at positions 246-248 was assigned as that coding for the initiator Met since it is the most 5'-proximal codon specifying a Met after the stop codon of TGA at positions 138-140. The "first ATG rule" reportedly holds for the vast majority of eucaryotic mRNAs (Kozak, 1984). The nucleotide sequence surrounding the tentative initiator Met codon is GACATGG. This differs somewhat from the optimal sequence of ACCATGG for translation initiation sites proposed by Kozak (1986). A purine is present, however, in a critical position located three nucleotides upstream of the ATG codon. The length of 5' untranslated regions in eucaryotic mRNAs can vary, with the majority (~70%) being in the range of 20-80 nucleotides (Kozak, 1984). The 245 nucleotides upstream from the apparent initiator Met represent a rather long

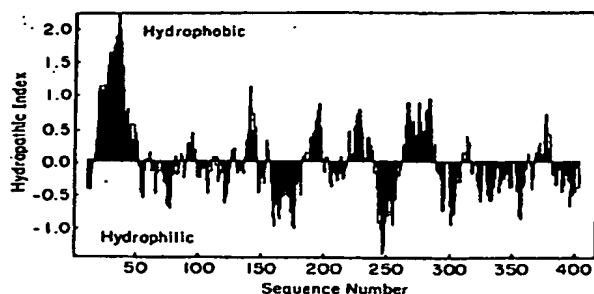


FIGURE 3: Hydropathy analysis of the deduced amino acid sequence of hepsin. The method of Kyte and Doolittle (1982) was employed, using a window of 20 residues. The peak spanning residues 18–44 represents the putative transmembrane domain.

5' untranslated region for hepsin. Although the precise role of the 5' untranslated sequence in mRNAs has not been established, it has been suggested that secondary structure(s) in long 5' untranslated regions may be involved in the regulation of transcription or translation (Kozak, 1984).

In contrast to most other serine proteases, the cDNA sequence coding for hepsin did not predict the presence of a typical signal peptide. However, hydropathy analysis (Kyte & Doolittle, 1982) revealed the presence of a single, very hydrophobic domain of 27 residues near the amino terminus of the molecule (residues 18–44, Figure 3). This hydrophobic domain, starting 18 residues downstream from the apparent initiator Met, contains no charged amino acids and is sufficiently long and nonpolar to span a lipid bilayer. Furthermore, this potential membrane-spanning domain is flanked on either side by charged amino acids, which may serve to help anchor the protein in a membrane.

From restriction enzyme mapping and DNA sequencing, it was found that clones HepG2UW17 and HepG2UW2 had additional sequences near their 5' ends that were not present in the other cDNA inserts. Beginning at position 192 in the nucleotide sequence, clone HepG2UW17 contained an additional 580 bp of DNA. This sequence was as follows: GTAAGGACAAGGGCCCCAGACTCACAGTTCCA-GCCCTGAGGACAGGGGTTCCCTCATCCCCCAC-CCAGCCTAATGCCACCTCCTAATAGAGGGGTT-CCTGGGACCTGAAGAGGGGGCACTATGACGT-CTCCCCAAGCACCTAGGTGTTCTGTCTCCTCT-TCCCTCAGACTCAGCCGTTGGACCCAGTCCTTT-CCTCCCCAGACCCAGGAGTTCAGCCCTCAGGC-CCCTCCTCCCTCATACTAGGGAGTCTGGCCCC-CAAATTCCTCCTTTCCCAAGACTTATGATTTCA-GGTCCTCAGCTGTCTCCTCCCTCAAACCGGGAT-CCTCAGTCCCCGTGCTCCACAGGCTCAGGCATG-GGGGTCCCCATCCCTGCAATCCAGGCGTCCCC-CCGCTGCTGGTCAGACACTGACCCCATCCTTGA-ACCCAGGCCAATCTGCGTCCGTGATCACGGCGT-GCTCTGGCCAAGGCCAGTCCCTACAGCCTGCC-TGGATGGACGCGCTGGGACTGGGGCGCCAGGA-CTGGGCTGGGCTGGGCTCCCCAGGCCCTGCT-CCCCGTCCATCTCCTCACAG. Analysis of this sequence suggests that this insertion probably represents an unspliced intron or a remnant of an intron. The underlined hexanucleotide sequences at the beginning and end of this sequence, GTAAGG and TCACAG, respectively, conform to consensus hexanucleotide sequences found at the 5' and 3' ends of introns adjacent to intron/exon splice junctions (Breathnach & Chambon, 1981; Nevins, 1983). The GTAAGG donor site and the TCACAG acceptor site are probably used for splic-

ing-out this intronic sequence in the majority of the mRNA molecules coding for hepsin. In the case of clone HepG2UW17, this sequence was not spliced-out when the mRNA molecule that gave rise to this particular insert was being processed. The additional sequence near the 5' end of clone HepG2UW2 is also probably due to improper splicing of the same intron. In this case, the cellular splicing apparatus apparently used the proper donor site (GTAAGG, underlined above), but an alternative acceptor site (ACCCAG, underlined above). This removed most of the intronic sequence but left behind 145 nucleotides. With the exception of these two probable splicing errors, no other differences were detected among the cDNA inserts in regions where overlapping sequences were obtained.

At the 3' end of the cDNA, the sequence of AATAAA was present 14 nucleotides upstream from the polyadenylation site. This sequence, which generally occurs 10–30 nucleotides upstream from the poly(A) tail, apparently functions as a signal for polyadenylation by either specifying the proper cleavage site of mRNA transcripts or serving as a recognition sequence for poly(A) polymerase (Proudfoot & Brownlee, 1976; Nevins, 1983).

The base composition of the cDNA coding for hepsin was particularly rich in G and C. The total nucleotide composition was calculated to be 17.0% A, 19.1% T, 31.2% G, and 32.5% C. The 245 bp 5' untranslated region contained an even higher content of C, and its base composition was calculated to be 17.1% A, 12.6% T, 28.5% G, and 41.6% C.

Besides the open reading frame that codes for hepsin, an unusually long open reading frame was observed in the inverted sequence of this cDNA. This open reading frame spanned 1353 nucleotides (nucleotides 105–1457 in the inverted sequence). The amino acid sequence deduced from this open reading frame was used in a search of the protein sequence database (National Biomedical Research Foundation, Washington, DC), but little significant sequence identity was found with any other known protein. Furthermore, there were no Met residues in the deduced amino acid sequence that could serve as a start site for translation.

DISCUSSION

Analysis of the cDNA sequence presented for hepsin indicates that it codes for a protein that is a member of the serine protease family. The cDNA coding for hepsin was isolated from cDNA libraries prepared from human liver and Hep G2 cell line mRNA. Preliminary data by Northern analysis indicate that the mRNA coding for hepsin is also expressed in a human osteosarcoma cell line. It is either not expressed or expressed only at very low levels in human endothelial cells, smooth muscle cells, and skin fibroblasts, as determined by Northern analysis.

The amino acid sequence of hepsin, deduced from the nucleotide sequence of its cDNA, is very similar to other serine proteases, especially in those regions that are highly conserved among this group of enzymes. It contains His, Asp, and Ser residues at positions 203, 257, and 353, respectively. These amino acids are analogous to the His₃₇, Asp₁₀₂, and Ser₁₉₅ residues in chymotrypsin that constitute the catalytic triad essential for enzymatic activity (Blow et al., 1969). The presence of an Asp (as opposed to a Ser) at position 347 suggests that hepsin possesses a substrate specificity similar to that of trypsin (Steitz et al., 1969; Hartley, 1970). This residue is thought to contribute to substrate binding in the active site of serine proteases and, for trypsin-like serine proteases, results in a preference for basic amino acids.

The cDNA sequence predicts an Arg-Ile-Val-Gly-Gly ac-

Rep:

Fact:

Pro:

Fact:

Fact:

FIGURE of fact 1982). the seq protein.

tivation hepsin to an peptid consist 1–162, and a carbox serine the no: expect: togeth (Natio: showe: serine

These among boxt-t shares in four 4). C: degree proteas

Whe differ: emerge occurri

(Hartle are muc analysis

et al. (1 six vari: conserv

lations a of the pr and act

whereas their ur compari

hepsin v apparent and vari

The b just prio sequence

X (Leyti Factor 2 cursors e

and rele: the activ

activation serine pr it seems j

Hepsin (119-154)	C V D E - G R L P R T Q R L L E V I S V - C D C P R G R F L A A I - - - C Q D - - - - C G
Factor X (89-133)	C S L D N G G D C D Q F C B E E Q N S V V - C S C A R G Y T L A D N G K A C I P T G P Y P C G
Protein C (98-142)	C S L D N G G G C T R Y C L E E V G W R R - C S C A P G Y K L G D D L L Q C H P A V K F P C G
Factor VII (91-136)	C V N E N G G C E Q Y C S D I T G T K R S C R C D E G Y S L L A D G V S C T P T V E Y P C G
Factor IX (88-133)	C N I K N G R C E Q P C K N S A D N E Y V C S C T E Q Y R L A E N Q K S C E P A V P P P C G

FIGURE 4: Comparison of the carboxyl-terminal end of the noncatalytic chain of hepsin with corresponding regions in the noncatalytic chains of factor X (McMullen et al., 1983), protein C (Foster & Davie, 1984), factor VII (Hagen et al., 1986), and factor IX (Kurachi & Davie, 1982). Gaps have been inserted to bring the protein sequences into better alignment. The numbers in parentheses refer to the location of the sequence in that particular protein. Amino acids are boxed if they are found at the same location in hepsin and one or more of the other proteins.

tivation site sequence (residues 162-166). This suggests that hepsin is synthesized as an inactive zymogen which is converted to an active serine protease by cleavage of the Arg₁₆₂-Ile₁₆₃ peptide bond. The resulting active serine protease would consist of two chains, including a noncatalytic chain (residues 1-162) derived from the amino-terminal end of the zymogen and a catalytic chain (residues 163-417) derived from the carboxyl-terminal end. By analogy with the various plasma serine proteases, the Cys residues at positions 153 and 277 in the noncatalytic and catalytic chains, respectively, could be expected to form a disulfide bond that holds the two chains together. A computer search of the protein sequence database (National Biomedical Research Foundation, Washington, DC) showed that a portion of hepsin differs substantially from all serine proteases for which there is sequence data available. These data also showed that the noncatalytic chain is unique among known protein sequences except for its extreme carboxyl-terminal region. This portion of the noncatalytic chain shares some sequence similarity with corresponding regions in four of the vitamin K dependent serine proteases (Figure 4). Conversely, the catalytic chain of hepsin exhibits a high degree of similarity with the catalytic chains of other serine proteases (Figure 5).

When the primary structures of the catalytic chains of different serine proteases are compared, the pattern that emerges is one of small stretches of highly similar sequence occurring at various intervals along the polypeptide chain (Hartley & Shotton, 1971). Furthermore, internal residues are much more highly conserved than external ones. In their analysis of the catalytic chains of several serine proteases, Furie et al. (1982) identified seven conserved regions separated by six variable regions. The variable regions, which show little conservation of sequence, in addition to containing short deletions and insertions, are thought to be located on the surface of the protein. This helps to explain why the internal structures and active sites of different serine proteases appear similar, whereas their surfaces, which play a major role in determining their unique substrate specificities, vary considerably. By comparing the amino acid sequence of the catalytic chain of hepsin with those of other serine proteases (Figure 5), it is apparent that hepsin also follows the same pattern of conserved and variable regions.

The highly basic sequence Arg-Arg-Lys (residues 155-157) just prior to the apparent activation site is similar to the basic sequences that also precede the activation sites in human factor X (Leytus et al., 1984) and protein C (Foster & Davie, 1984). Factor X and protein C are synthesized as single-chain precursors and are converted to two-chain zymogens by cleavage and release of these basic residues. Subsequent cleavages at the activation sites for factor X and protein C release short activation peptides and result in the generation of an active serine protease. If the analogy is extended to include hepsin, it seems possible that this protein may also exist as a two-chain

zymogen that releases a short peptide (e.g., Leu-Pro-Val-Asp-Arg) upon its conversion to an active enzyme.

Compared with other serine proteases, the number and positions of 9 out of the 10 cysteine residues in the catalytic chain of hepsin are highly conserved. On the basis of the known disulfide bridge arrangement in chymotrypsin (Keil et al., 1963; Brown & Hartley, 1966), trypsin (Kauffman, 1965), prothrombin (Magnusson et al., 1975), plasmin (Sottrup-Jensen et al., 1978; Wiman, 1977), and factor X (Hojrup & Magnusson, 1987), and by analogy with other serine proteases, four intrachain disulfide bonds at cysteine pairs 188/204, 291/359, 322/338, and 349/381 would be expected. In addition, Cys₂₇₇ is probably involved in a disulfide linkage with the noncatalytic chain. The remaining Cys₃₇₂ has no analogous counterpart in other serine proteases. One possibility is that this extra Cys may participate in an interchain disulfide bridge between two monomers of hepsin, analogous to that proposed for factor XI (Fujikawa et al., 1986). In the noncatalytic chain of hepsin, the cDNA sequence predicts the presence of nine Cys residues. Cys₁₅₃ is probably involved in the disulfide linkage with the catalytic chain. This leaves an even number of Cys residues in the noncatalytic chain that could form intrachain disulfide bonds.

From crystallographic and kinetic studies of chymotrypsin and trypsin and from knowledge of their primary structures, it has been possible to identify residues in these enzymes that are involved in substrate binding and catalysis [reviewed in Birktoft et al. (1970), Hartley and Shotton (1971), and Kraut (1977)]. Since some of these residues are essential for proper function, it was of interest to make a more detailed comparison with hepsin (Figure 5) and to determine whether hepsin possessed these same essential residues.

(a) During the conversion of chymotrypsinogen to chymotrypsin, the peptide backbone of segment 187-193 becomes more extended, resulting in the creation of a substrate binding pocket (Kraut, 1971). The peptide backbone of residues Ser₁₈₉-Ser₁₉₀-Cys₁₉₁-Met₁₉₂ forms one side of this substrate binding pocket in chymotrypsin (Steitz et al., 1969). This sequence is Asp₁₈₉-Ser₁₉₀-Cys₁₉₁-Gln₁₉₂ in trypsin and Asp₁₈₉-Ala₁₉₀-Cys₁₉₁-Gln₁₉₂ in hepsin.

(b) The opposite side of the substrate binding pocket in chymotrypsin is lined by residues Ser₂₁₄-Trp₂₁₅-Gly₂₁₆. The peptide backbone of these residues is thought to interact with the side chains of the substrate for properly orienting the bond that is to be cleaved (Steitz et al., 1969). This stretch of amino acids is also present in hepsin.

(c) Hydrogen bonding between Cys₁₉₁/Asp₁₉₄ and Asp₁₉₄/Gly₁₉₇ provides a rigid structure in the peptide backbone of chymotrypsin in the vicinity of the active site. This helps to hold the active-site Ser₁₉₅ in the proper orientation and is maintained only if Gly residues are present at positions 193 and 196 (Birktoft et al., 1970). Hepsin also has Gly residues at these two positions.

side cl
(1970)
189 m
acidic

(c)

chains
chymo
or cha
pocket
In hep
226 is

(f)

a flexil
in chy
al., 19
nonpol
trypsin
hepsin,

(g)

in serin
tivation
also pr

The

a poten
to sever
recepto
et al., 1

(van D)

protein:

that is

hydropl

though

drophol

membr

terminu

facing i

& Dricl

1985; S

brane-sj

membra

the mec

of the er

the amio

terminu

sequence

in hepsi

membra

of trans

would p

hepsin v

processe

involve l

well char

serine p

these pr

It is di

of hepsi

it probab

fibrinoly

expresse

thesize a

involved

synthesiz

propeptic

depende

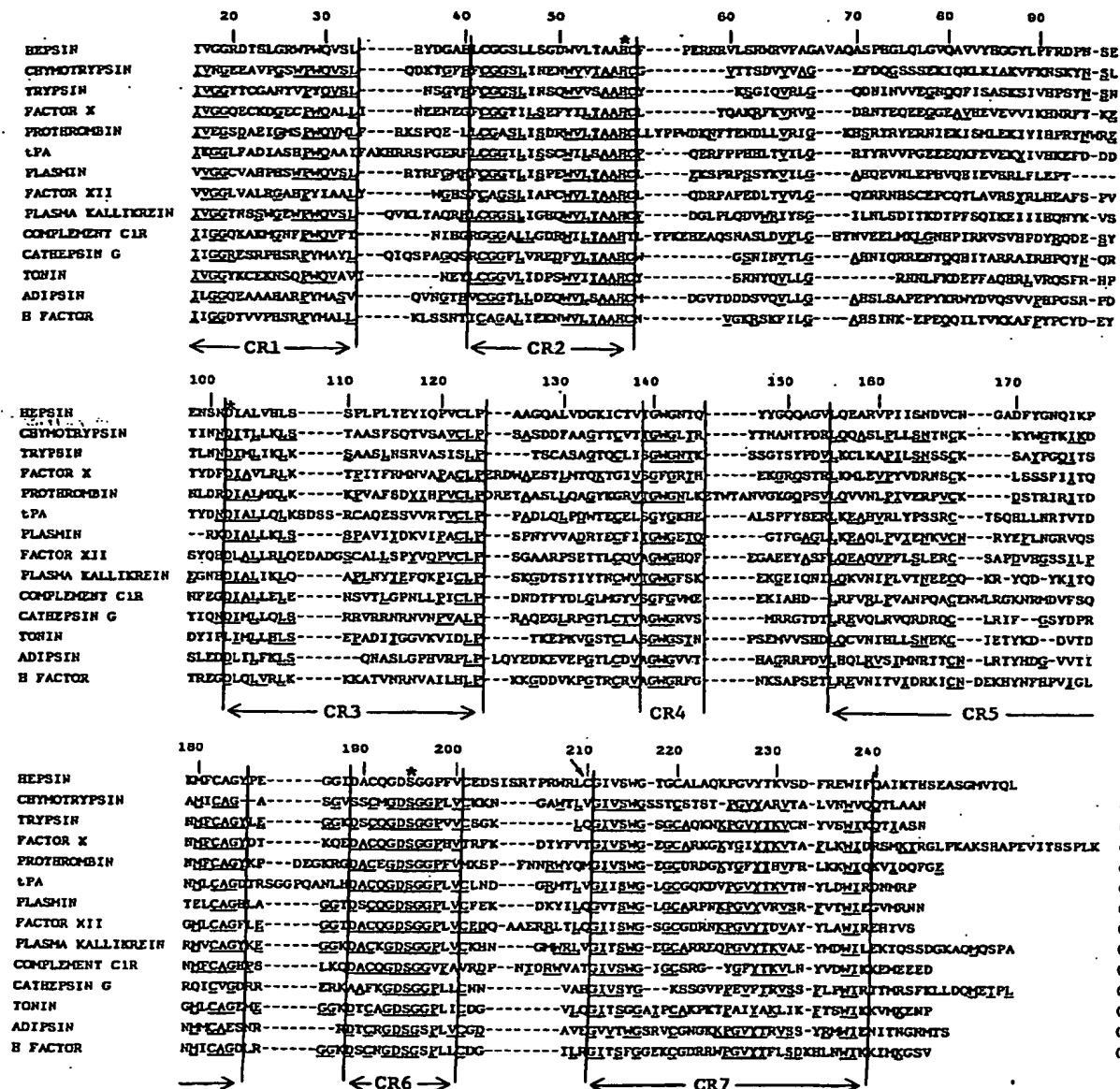


FIGURE 5: Comparison of the presumed catalytic chain of hepsin with the catalytic chains of a variety of other serine proteases, including bovine chymotrypsin (Hartley, 1964; Meloun et al., 1966; Hartley & Kauffman, 1966; Blow et al., 1969), bovine trypsin (Walsh & Neurath, 1964; Mikes et al., 1966; Eyl & Inagami, 1970; Hartley, 1970), human factor X (Leytus et al., 1986b), human prothrombin (Degen et al., 1983), human tissue plasminogen activator (tPA) (Pennica et al., 1983), human plasmin (Malinowski et al., 1984), human factor XII (Fujikawa & McMullen, 1983), human plasma kallikrein (Chung et al., 1986), human complement C1r (Arlaud & Gagnon, 1983), human cathepsin G (Salvesen et al., 1987), rat submaxillary tonin (Lazure et al., 1984), mouse adipsin (Cook et al., 1985), and mouse H factor (Gershenfeld & Weissman, 1986). In this figure, the numbering of residues follows the standard chymotrypsinogen notation (Hartley, 1970), and the boundaries of seven conserved regions (CR1-7) are essentially the same as those designated by Furie et al. (1982). Since variable regions show minimal sequence conservation, little attempt was made to optimize the homology in these regions. Otherwise, gaps have been inserted to bring the sequences into better alignment. Asterisks have been placed above the active-site residues His₅₇, Asp₁₀₂, and Ser₁₉₅ that compose the catalytic triad. An arrow indicates the location of the extra Cys residue in the sequence of hepsin. Residues are underlined when the same amino acid is found at the same position in hepsin. The percentage listed in parentheses at the end of each sequence represents the extent of similarity between hepsin and that protein, as calculated from this alignment.

(d) All acidic (Asp and Glu) and basic (Arg, Lys, and His) side chains are placed on the surface of chymotrypsin, with the exception of Asp₁₀₂ and Asp₁₉₄, which are buried in the interior of the molecule. In trypsin, there is an additional buried acidic side chain at Asp₁₈₉. Hepsin contains the two

buried Asp residues common to both chymotrypsin and trypsin, namely, Asp₁₀₂ and Asp₁₉₄. In addition, at the position which has the greatest influence on substrate specificity (position 189), hepsin contains an Asp residue. Thus, it is predicted that hepsin would have a preference for substrates with basic

side chains. It is of interest to note that Shotton and Watson (1970) made the prediction that a basic residue at position 189 might result in a serine protease with a preference for acidic side chains.

(e) In the three-dimensional model for elastase, the side chains of Val₂₁₆ and Thr₂₂₆, replacing Gly₂₁₆ and Gly₂₂₆ in chymotrypsin and trypsin, block the entrance of hydrophobic or charged substrates with bulky side chains from the binding pocket (Shotton & Hartley, 1970; Shotton & Watson, 1970). In hepsin, the presence of Gly residues at positions 216 and 226 is preserved.

(f) The side chain of residue 192 has been described as being a flexible cover to the entrance of the substrate binding pocket in chymotrypsin (Steitz et al., 1969) and trypsin (Krieger et al., 1974). In chymotrypsin, Met₁₉₂ may help provide a nonpolar environment for substrate side chains, whereas in trypsin Gln₁₉₂ may provide a more polar environment. In hepsin, position 192 is Gln.

(g) The sequence Gly₁₄₀-Trp₁₄₁-Gly₁₄₂ is highly conserved in serine proteases and is presumed to be involved in the activation process (Fehlhammer et al., 1977). This sequence is also present in hepsin.

The absence of a typical signal peptide and the presence of a potential transmembrane domain in hepsin are analogous to several other proteins recently described. Asialoglycoprotein receptor (Holland et al., 1984), transferrin receptor (Schneider et al., 1984), and plasma cell membrane glycoprotein PC-1 (van Driel & Goding, 1987) are examples of transmembrane proteins which lack a typical amino-terminal signal peptide that is cleaved during biosynthesis. These proteins possess hydrophobic domains near their amino termini which are thought to function as internal signal sequences. The hydrophobic domains direct insertion of these proteins into the membrane of the endoplasmic reticulum, leaving the amino terminus facing the cytoplasm and the carboxyl terminus facing into the lumen of the endoplasmic reticulum (Holland & Drickamer, 1986; Zerial et al., 1986; Wickner & Lodish, 1985; Spiess & Lodish, 1986). If a protein with a membrane-spanning domain is ultimately destined for the plasma membrane, its orientation at the cell surface is determined by the mechanism by which it was inserted into the membrane of the endoplasmic reticulum. For the cases mentioned above, the amino terminus faces the cytoplasm, whereas the carboxyl terminus is extracellular. The lack of an amino-terminal signal sequence and the presence of an internal hydrophobic domain in hepsin suggest that it is synthesized and integrated into membranes in a manner similar to the above-mentioned group of transmembrane proteins. If this were the case, then one would predict that the carboxyl-terminal catalytic chain of hepsin would be on the outside of the cell. There are many processes occurring extracellularly near the cell surface that involve limited proteolysis. Although these have not yet been well characterized, an activatable, trypsin-like, transmembrane serine protease may be an important participant in some of these processes.

It is difficult to speculate as to the true physiological function of hepsin. Since it may be a membrane-associated protein, it probably is not participating in such processes as coagulation, fibrinolysis, complement activation, etc., unless it is also being expressed by endothelial or blood cells. Since liver cells synthesize and secrete many different proteins, hepsin might be involved in the modification of other proteins as they are being synthesized or secreted. This could include the removal of propeptides from hormones, growth factors, or the vitamin K dependent proteases or the activation or inactivation of other

proteins. It is unclear, however, how hepsin is converted from a zymogen to an active enzyme and whether this involves another serine protease or whether hepsin is capable of autoactivation. Answers to these questions will require additional experimentation.

ACKNOWLEDGMENTS

We thank Drs. Akitada Ichinose, Jose Lopez, Kazuo Fujikawa, and Dominic Chung for valuable discussions and advice. We also thank Lois Swenson for her assistance in the preparation of the manuscript.

REFERENCES

- Arlaud, G. J., & Gagnon, J. (1983) *Biochemistry* 22, 1758-1764.
- Benton, W. D., & Davis, R. W. (1977) *Science (Washington, D.C.)* 196, 180-182.
- Biggin, M. D., Gibson, T. J., & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 3963-3965.
- Billings, P. C., Carew, J. A., Keller-McGandy, C. E., Goldberg, A. L., & Kennedy, A. R. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 4801-4805.
- Birktoft, J. J., Blow, D. M., Henderson, R., & Steitz, T. A. (1970) *Philos. Trans. R. Soc. London, B* 257, 67-76.
- Birnboim, H. C., & Doly, J. (1979) *Nucleic Acids Res.* 7, 1513-1523.
- Blow, D. M., Birktoft, J. J., & Hartley, B. S. (1969) *Nature (London)* 221, 337-340.
- Breathnach, R., & Chambon, P. (1981) *Annu. Rev. Biochem.* 50, 349-383.
- Brown, J. R., & Hartley, B. S. (1966) *Biochem. J.* 101, 214-228.
- Chandra, T., Stackhouse, R., Kidd, V. J., & Woo, S. L. C. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 1845-1848.
- Christman, J. K., Silverstein, S. C., & Acs, G. (1977) in *Proteinases in Mammalian Cells and Tissues* (Barrett, A. J., Ed.) pp 91-149, Elsevier, Amsterdam and New York.
- Chung, D. W., Fujikawa, K., McMullen, B. A., & Davie, E. W. (1986) *Biochemistry* 25, 2410-2417.
- Collen, D. (1980) *Thromb. Haemostasis* 43, 77-89.
- Cook, K. S., Groves, D. L., Min, H. Y., & Spiegelman, B. M. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 6480-6484.
- Cook, K. S., Min, H. Y., Johnson, D., Chaplinsky, R. J., Flier, J. S., Hunt, C. R., & Spiegelman, B. M. (1987) *Science (Washington, D.C.)* 237, 402-405.
- Cromlish, J. A., Seidah, N. G., & Chretien, M. (1986) *J. Biol. Chem.* 261, 10850-10858.
- Davie, E. W., Fujikawa, K., Kurachi, K., & Kisiel, W. (1979) *Adv. Enzymol. Relat. Areas Mol. Biol.* 48, 277-318.
- Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., Ed.) Vol. 5, Suppl. 3, pp 1-8, National Biomedical Research Foundation, Washington, DC.
- Dayhoff, M. O., Barker, W. C., & Hunt, L. T. (1983) *Methods Enzymol.* 91, 524-545.
- Degen, S. J. F., MacGillivray, R. T. A., & Davie, E. W. (1983) *Biochemistry* 22, 2087-2097.
- Eisenbauer, D. A., & McDonald, J. K. (1986) *J. Biol. Chem.* 261, 8859-8865.
- Eyl, A., & Inagami, T. (1970) *Biochem. Biophys. Res. Commun.* 38, 149-155.
- Fehlhammer, H., Bode, W., & Huber, R. (1977) *J. Mol. Biol.* 111, 415-438.
- Foster, D., & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 4766-4770.

- Fujikawa, K., & McMullen, B. A. (1983) *J. Biol. Chem.* 258, 10924-10933.
- Fujikawa, K., Chung, D. W., Hendrickson, L. E., & Davie, E. W. (1986) *Biochemistry* 25, 2417-2424.
- Furie, B., Bing, D. H., Feldmann, R. J., Robison, D. J., Burnier, J. P., & Furie, B. C. (1982) *J. Biol. Chem.* 257, 3875-3882.
- Gergen, J. P., Stern, R. H., & Wensink, P. C. (1979) *Nucleic Acids Res.* 7, 2115-2136.
- Gershensfeld, H. K., & Weissman, I. L. (1986) *Science (Washington, D.C.)* 232, 854-858.
- Hagen, F. S., Gray, C. L., O'Hara, P., Grant, F. J., Saari, G. C., Woodbury, R. G., Hart, C. E., Insley, M., Kisiel, W., Kurachi, K., & Davie, E. W. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 2412-2416.
- Hartley, B. S. (1964) *Nature (London)* 201, 1284-1287.
- Hartley, B. S. (1970) *Philos. Trans. R. Soc. London, B* 257, 77-87.
- Hartley, B. S., & Kauffman, D. L. (1966) *Biochem. J.* 101, 229-231.
- Hartley, B. S., & Shotton, D. M. (1971) *Enzymes (3rd Ed.)* 3, 323-373.
- Hojrup, P., & Magnusson, S. (1987) *Biochem. J.* 245, 887-892.
- Holland, E. C., & Drickamer, K. (1986) *J. Biol. Chem.* 261, 1286-1292.
- Holland, E. C., Leung, J. O., & Drickamer, K. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 7338-7342.
- Kauffman, D. L. (1965) *J. Mol. Biol.* 12, 929-932.
- Keil, B., Prusik, Z., & Sorm, F. (1963) *Biochim. Biophys. Acta* 78, 559-561.
- Kozak, M. (1984) *Nucleic Acids Res.* 12, 857-872.
- Kozak, M. (1986) *Cell (Cambridge, Mass.)* 44, 283-292.
- Kraut, J. (1971) *Enzymes (3rd Ed.)* 3, 165-183.
- Kraut, J. (1977) *Annu. Rev. Biochem.* 46, 331-358.
- Krieger, M., Kay, L. M., & Stroud, R. M. (1974) *J. Mol. Biol.* 83, 209-230.
- Kurachi, K., & Davie, E. W. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 6461-6464.
- Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol.* 157, 105-132.
- LaBombardi, V. J., Shaw, E., DiStefano, J. F., Beck, G., Brown, F., & Zucker, S. (1983) *Biochem. J.* 211, 695-700.
- Lazure, C., Leduc, R., Seidah, N. G., Thibault, G., Genest, J., & Chretien, M. (1984) *Nature (London)* 307, 555-558.
- Leytus, S. P., Chung, D. W., Kisiel, W., Kurachi, K., & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 3699-3702.
- Leytus, S. P., Kurachi, K., Sakariassen, K. S., & Davie, E. W. (1986a) *Biochemistry* 25, 4855-4863.
- Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. W. (1986b) *Biochemistry* 25, 5098-5102.
- Lobe, C. G., Finlay, B. B., Paranchych, W., Paetkau, V. H., & Bleackley, R. C. (1986) *Science (Washington, D.C.)* 232, 858-861.
- Lundgren, S., Ronne, H., Rask, L., & Peterson, P. A. (1984) *J. Biol. Chem.* 259, 7780-7784.
- Magnusson, S., Petersen, T. E., Sottrup-Jensen, L., & Claeys, H. (1975) in *Proteases and Biological Control* (Reich, E., Rifkin, D. B., & Shaw, E., Eds.) pp 123-249, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Malinowski, D. P., Sadler, J. E., & Davie, E. W. (1984) *Biochemistry* 23, 4243-4250.
- Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Masson, D., & Tschopp, J. (1987) *Cell (Cambridge, Mass.)* 49, 679-685.
- Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol.* 65, 499-560.
- McMullen, B. A., Fujikawa, K., Kisiel, W., Sasagawa, T., Howald, W. N., Kwa, E. Y., & Weinstein, B. (1983) *Biochemistry* 22, 2875-2884.
- Meloun, B., Kluh, I., Kostka, V., Moravek, L., Prusik, Z., Vanecek, J., Keil, B., & Sorm, F. (1966) *Biochim. Biophys. Acta* 130, 543-546.
- Messing, J. (1983) *Methods Enzymol.* 101, 20-78.
- Micard, D., Sobrier, M. L., Couderc, J. L., & Dastugue, B. (1985) *Arial. Biochem.* 148, 121-126.
- Mikes, O., Holeysovsky, V., Tomasek, V., & Sorm, F. (1966) *Biochem. Biophys. Res. Commun.* 24, 346-352.
- Neurath, H., & Walsh, K. A. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 3825-3832.
- Nevins, J. R. (1983) *Annu. Rev. Biochem.* 52, 441-446.
- Pasternack, M. S., Verret, C. R., Liu, M. A., & Eisen, H. N. (1986) *Nature (London)* 322, 740-743.
- Pennica, D., Holmes, W. E., Kohr, W. J., Harkins, R. N., Vehar, G. A., Ward, C. A., Bennett, W. F., Yelverton, E., Seeburg, P. H., Heyneker, H. L., & Goeddel, D. V. (1983) *Nature (London)* 301, 214-221.
- Proudfoot, N. J., & Brownlee, G. G. (1976) *Nature (London)* 263, 211-214.
- Reid, K. B. M., & Porter, R. R. (1981) *Annu. Rev. Biochem.* 50, 433-464.
- Salvesen, G., Farley, D., Shuman, J., Przybyla, A., Reilly, C., & Travis, J. (1987) *Biochemistry* 26, 2289-2293.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463-5467.
- Schneider, C., Owen, M. J., Banville, D., & Williams, J. G. (1984) *Nature (London)* 311, 675-678.
- Shotton, D. M., & Hartley, B. S. (1970) *Nature (London)* 225, 802-806.
- Shotton, D. M., & Watson, H. C. (1970) *Nature (London)* 225, 811-816.
- Sottrup-Jensen, L., Claeys, H., Zajdel, M., Petersen, T. E., & Magnusson, S. (1978) *Prog. Chem. Fibrinolysis Thrombolysis* 3, 191-209.
- Spies, M., & Lodish, H. F. (1986) *Cell (Cambridge, Mass.)* 44, 177-185.
- Steitz, T. A., Henderson, R., & Blow, D. M. (1969) *J. Mol. Biol.* 46, 337-348.
- Tanaka, K., Nakamura, T., & Ichihara, A. (1986) *J. Biol. Chem.* 261, 2610-2615.
- van Driel, I. R., & Goding, J. W. (1987) *J. Biol. Chem.* 262, 4882-4887.
- Vieira, J., & Messing, J. (1982) *Gene* 19, 259-268.
- Walsh, K. A., & Neurath, H. (1964) *Proc. Natl. Acad. Sci. U.S.A.* 52, 884-889.
- Watt, K. W. K., Lee, P.-J., M'Timkulu, T., Chan, W.-P., & Loo, R. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 3166-3170.
- Wickner, W. T., & Lodish, H. F. (1985) *Science (Washington, D.C.)* 230, 400-407.
- Wiman, B. (1977) *Eur. J. Biochem.* 76, 129-137.
- Young, J. D.-E., Leong, L. G., Liu, C.-C., Damiano, A., Wall, D. A., & Cohn, Z. A. (1986) *Cell (Cambridge, Mass.)* 47, 183-194.
- Zerial, M., Melancon, P., Schneider, C., & Garoff, H. (1986) *EMBO J.* 5, 1543-1550.

Sp

T
Tran
copro
iron i
additi
bound
Such
the pr
(1986
know
large
f elec
trast-
Beacu
probe
Gd(II
The
relati
sequen
with c
of the
the si
spectr
(Bald
that d
to the
spectr
of Gd
EXPE
Ma
from
or pur
Th
tional
Co
ment o

Exhibit 20

Molecular Cloning of cDNA for Matriptase, a Matrix-degrading Serine Protease with Trypsin-like Activity*

(Received for publication, November 23, 1998, and in revised form, April 8, 1999)

Chen-Yong Lin, Joanna Anders, Michael Johnson, Qingxiang Amy Sang‡, and Robert B. Dickson§

From the Lombardi Cancer Center, Georgetown University Medical Center, Washington, D. C. 20007

A major protease from human breast cancer cells was previously detected by gelatin zymography and proposed to play a role in breast cancer invasion and metastasis. To structurally characterize the enzyme, we isolated a cDNA encoding the protease. Analysis of the cDNA reveals three sequence motifs: a carboxyl-terminal region with similarity to the trypsin-like serine proteases, four tandem cysteine-rich repeats homologous to the low density lipoprotein receptor, and two copies of tandem repeats originally found in the complement sub-components C1r and C1s. By comparison with other serine proteases, the active-site triad was identified as His-484, Asp-539, and Ser-633. The protease contains a characteristic Arg-Val-Val-Gly-Gly motif that may serve as a proteolytic activation site. The bottom of the substrate specificity pocket was identified to be Asp-627 by comparison with other trypsin-like serine proteases. In addition, this protease exhibits trypsin-like activity as defined by cleavage of synthetic substrates with Arg or Lys as the P1 site. Thus, the protease is a mosaic protein with broad spectrum cleavage activity and two potential regulatory modules. Given its ability to degrade extracellular matrix and its trypsin-like activity, the name matriptase is proposed for the protease.

Elevated proteolytic activity has been implicated in neoplastic progression. Although the exact role(s) of proteolytic enzymes in the progression of tumor remains unclear, it seems that proteases may be involved in almost every step of the development and spread of cancer. A widely proposed view is that proteases contribute to the degradation of extracellular matrix and to tissue remodeling and are necessary for cancer invasion and metastasis. A wide array of extracellular matrix-degrading proteases have been discovered, the expression of some of which correlates with tumor progression, as reviewed by Magnatti and Rifkin (1). The plasmin/urokinase-type plas-

minogen activator system and the 72-kDa gelatinase (MMP-2)/membrane-type MMP system have received the most attention for their potential roles in the process of invasion of breast cancer and other carcinomas. However, both systems appear to be largely synthesized by stromal cells *in vivo* (2–5) and require indirect mechanisms for their recruitment and activation on the surfaces of cancer cells. The stromal origins of these well characterized extracellular matrix-degrading proteases may suggest that cancer invasion is an event that either depends entirely upon stromal-epithelial cooperation or is controlled by some other unknown epithelium-derived protease(s). A search for these epithelium-derived proteolytic systems that may interact with the plasmin/urokinase-type plasminogen activator system and/or with the MMP family could provide a missing link in our understanding of malignant invasion.

We have pursued studies of a novel protease with the hypothesis that a tumor itself may be a major source of proteases important for multiple aspects of malignant behavior, including invasion and metastasis. To this end, we systematically altered several conditions such as the pH using gelatin zymography to search for potentially important breast cancer cell-derived gelatinases. This search led us to the discovery of a major protease, which on a gelatin zymogram had a slightly alkaline pH optimum and a size between those of MMP-2 and MMP-9 in T-47D human breast cancer cells (6). We now propose to call this protease matriptase. Matriptase has been purified from T-47D cell-conditioned medium and has been used as an immunogen to produce monoclonal antibodies (7). Although matriptase was initially isolated from cell-conditioned medium, three lines of evidence, including immunofluorescence staining, surface biotinylation, and subcellular fractionation, suggested that a portion of the enzyme molecules were localized on the surfaces of cells. Given its extracellular matrix-degrading activity and presentation on the surfaces of breast cancer cells, we hypothesize that matriptase may be involved in breast cancer invasion. To further characterize the newly discovered matrix-degrading protease in this study, we have purified the enzyme and its binding protein from human milk, a biological source of relatively high abundance. A cDNA clone for matriptase has now been generated and characterized.

MATERIALS AND METHODS

Cell Lines and Culture Conditions—COS-7 cells were maintained in modified Iscove's minimal essential medium (Biofluids, Inc., Rockville, MD) supplemented with 5% fetal calf serum (Life Technologies, Inc.).

Purification of Matriptase—To obtain enough matriptase for amino acid sequencing, the enzyme was isolated from human milk (39). Briefly, human milk from the Georgetown University Medical Center Milk Bank was precipitated and collected by addition of ammonium sulfate between 40 and 60% saturation. Matriptase was purified by a combination of CM-Sepharose and immunoaffinity chromatography.

Amino Acid Sequence Analysis—To obtain internal amino acid sequences, purified matriptase was separated by SDS-polyacrylamide gel electrophoresis and lightly stained with Coomassie Blue, and protein

* This work was supported by Specialized Program of Research Excellence in Breast Cancer Grant 1P50CA58158 (to R. B. D.) from the National Institutes of Health and by the Elsa U. Pardee Foundation (to Q. A. S.). Work performed at the Lombardi Cancer Center Macromolecular Synthesis and Sequencing Shared Resource was supported by National Institutes of Health Grant P30-CA51008. Cells and reagents were obtained from the Lombardi Cancer Center Tissue Culture Resource, supported by National Institutes of Health Grant P30-CA51005. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AF118224.

‡ Present address: Dept. of Chemistry, Florida State University, Tallahassee, FL 32306.

§ To whom correspondence and reprint requests should be addressed: Lombardi Cancer Center, Georgetown University Medical Center, Washington, D. C. 20007. Tel.: 202-687-4304; Fax: 202-687-7505.

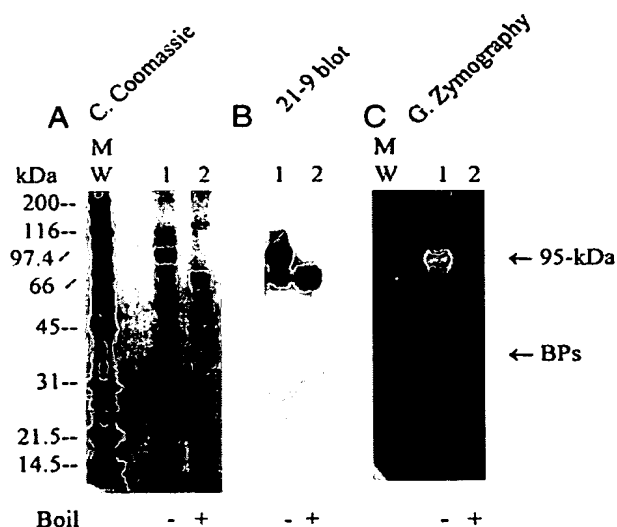


FIG. 1. Purification of matriptase in its 95-kDa complexed form from human milk. The partially purified 95-kDa matriptase complex from ion-exchange chromatography was loaded onto a mAb 21-9-Sepharose column. The bound proteins were eluted by glycine buffer, pH 2.4, and neutralized by addition of 2 M Trizma. The eluted proteins were incubated in 1× SDS sample buffer in the absence of reducing agents at room temperature (lanes 1; -Boil) or at 95 °C (lanes 2; +Boil) for 5 min. The samples were resolved by SDS-polyacrylamide gel electrophoresis and either stained by colloidal Coomassie (A) or subjected to immunoblot analysis using mAb 21-9 (B) or gelatin zymography (C). The 95-kDa matriptase complex was eluted from this affinity column as the major protein (A, lane 1); it was recognized by mAb 21-9 (B, lane 1); and it also exhibited gelatinolytic activity (C, lane 1). The 95-kDa matriptase complex was converted to matriptase by boiling (A, lane 2). The gelatinolytic activity of the 95-kDa protease was destroyed by boiling, but a low level of the gelatinolytic activity was survived and converted to matriptase (C, lane 2). A low level of uncomplexed matriptase was copurified with the 95-kDa matriptase complex by affinity chromatography (A, lane 1); it also exhibited gelatinolytic activity (C, lane 1). Immunoblot analysis enhanced the signal of the uncomplexed matriptase and reconfirmed its existence (B, lane 1). Several other polypeptides were also seen (A, lanes 1 and 2). Some of them could be the degraded products of the protease since they were recognized by mAb 21-9 after longer exposure to the x-ray film. A 40-kDa protein doublet was seen in low levels in a nonboiled sample (A, lane 1), but its levels were increased after boiling (A, lane 2). This 40-kDa doublet was not recognized by mAb 21-9 (B). We propose that these two polypeptides could be binding proteins (BPs) of matriptase. The sizes of the molecular mass markers are indicated.

bands were excised. Matriptase was then subjected to in-gel digestion and amino acid sequencing at the Howard Hughes Medical Institute Biopolymer Laboratory and W. M. Keck Foundation Biotechnology Resource Laboratory at Yale University. The amino-terminal sequences were determined as described previously (8). Briefly, the proteins were resolved by SDS-polyacrylamide gel electrophoresis, transferred to polyvinylidene difluoride membrane, and lightly stained with Coomassie Blue. The proteins were then excised and subjected to amino-terminal sequencing in the Chemistry Department of Florida State University (Tallahassee, FL). The two short sequences obtained were identical to a deduced amino acid sequence from a cDNA termed SNC19 (GenBank™ accession number U20428).

Amplification of an SNC19 cDNA from T-47D Breast Cancer Cells—An SNC19 cDNA clone was generated by reverse transcriptase-polymerase chain reaction utilizing mRNA from T-47D human breast cancer cells. Primer sequences for SNC19 (5'-CCTCCTCTTGGTCTT-GCTGGGG-3' and 5'-AGACCCGTCTGTTTCCAGG-3') were derived from the published sequence. Standard reverse transcription-polymerase chain reaction was conducted using the Advantage RT-PCR kit (CLONTECH). Products were analyzed on a 0.8% agarose gel; and the resultant band of ~2.8 kilobase pairs, corresponding to the expected product size, was excised from the gel, purified, and ligated into pCR2.1 (Invitrogen, San Diego, CA) by TA cloning (pCR-SNC19).

Sequencing—DNA sequencing was performed on an Applied Biosys-



FIG. 2. Western blot analysis of SNC19 protein expressed in COS cells using anti-matriptase mAb M32. The SNC19 fragment generated by reverse transcriptase-polymerase chain reaction was inserted into the expression vector pcDNA3.1 and transfected into COS-7 cells. Cell lysates from SNC19-transfected COS-7 cells (lane 1) and control COS-7 cells (lane 2) and the conditioned medium of T-47D human breast cancer cells (lane 3) were subjected to Western blot analysis using anti-matriptase mAb M32.

tems automated 377 DNA sequencer using standard methods, with the assistance of the Lombardi Cancer Center Sequencing and Synthesis Shared Resource. The sequences were assembled and analyzed with Lasergene software for Windows (DNASTAR, Inc., Madison, WI). The predicted protein sequence was compared with sequences in the Swiss-Prot data base at the National Center for Biotechnology Information using the BLAST network server.

Expression of SNC19 in COS-7 Cells—To verify that SNC19 encodes the matriptase cDNA, we constructed a eukaryotic expression vector (pcDNA/SNC19) utilizing the commercially available pcDNA3.1 vector (Invitrogen, San Diego, CA). A 2.83-kilobase pair *EcoRI* fragment containing the SNC19 cDNA was produced by digestion of pCR-SNC19 and cloned into the *EcoRI* site of pcDNA3.1. This construct contains the open reading frame of SNC19 driven by the cytomegalovirus promoter. Correct insertion of the SNC19 cDNA was verified by restriction mapping (data not shown). Transfections were carried out using SuperFect transfection reagent (QIAGEN Inc., Valencia, CA) as specified in the manufacturer's handbook. After 48 h, the matriptase-transfected COS-7 cells and the control COS-7 cells, which were transfected with LacZ to monitor transfection efficiency, were extracted with 1% Triton X-100 in 20 mM Tris-HCl, pH 7.4.

Immunoblot Analysis—Immunoblotting was conducted as described previously (7). Proteins were separated by 10% SDS-polyacrylamide gel electrophoresis, transferred to polyvinylidene fluoride membrane, and subsequently probed with anti-matriptase mAb¹ M32. Immunoreactive polypeptides were visualized using peroxidase-labeled secondary antiserum and the ECL detection system (Amersham Pharmacia Biotech).

Gelatin Zymography—Gelatin zymography was carried out as described previously with some modifications (13). Gelatin (1 mg/ml) as a substrate was copolymerized with regular SDS-polyacrylamide gel. Electrophoresis was performed at a constant current of 15 mA. The gelatin gels were washed three times with phosphate-buffered saline containing 2% Triton X-100 and incubated in phosphate-buffered saline at 37 °C overnight.

Cleavage of Synthetic Substrates—To demonstrate the trypsin-like activity of matriptase, various synthetic fluorescent protease substrates with arginine or lysine as the P1 site were tested with purified matriptase from human milk. Matriptase was assayed in 20 mM Tris buffer, pH 8.5, at 25 °C in a volume of 190 μl prior to addition of 10 μl of 2 mM substrate solution (to a final concentration of 0.1 mM). These substrates included *t*-butyloxycarbonyl (Boc)-Gln-Ala-Arg-7-amino-4-methylcoumarin (AMC), Boc-benzyl-Glu-Gly-Arg-AMC, Boc-Leu-Gly-Arg-AMC, Boc-benzyl-Asp-Pro-Arg-AMC, Boc-Phe-Ser-Arg-AMC, Boc-Val-Pro-Arg-AMC, succinyl-Ala-Phe-Lys-AMC, Boc-Leu-Arg-Arg-AMC, Boc-Gly-Lys-Arg-AMC, and Boc-Leu-Ser-Thr-Arg-AMC. These sub-

¹ The abbreviations used are: mAb, monoclonal antibody; Boc, *t*-butyloxycarbonyl; AMC, 7-amino-4-methylcoumarin; LDL, low density lipoprotein.

-357 CGCTGGGTGGTGTGCTGCGACGGGCGTGTGATCGGCTCTCTTGGCTTGGTGGGATCGGCTTCTGGTGGGATTTGCAATACACGG
-270 GACCTCGGTGTCTCAGAAGCTCTTCAATGGCTACATCAGGATCACCAATAGAAATTTTGTTGATGCTCAGCAACAATCCAACTCCATCGAG
-180 TTGTGAAGCTGGCTGCGAAGGTGAAGCAGCGCTGAAGCTGCTGTACAGCGGAGTCCCACTTCTGGGCGCTTCCACAAGGACGAGCTCGGCT
-90 GTGACGGCTCTCAGCGAGGCGAGCTCATCGCTTACTTGGTGTGATCTCAGCATCCCGAGCACTGGTGGAGGAGGCGGAGCGGCT
1 ATGGCGACGACGCGGTAGTCTAGTCTGCTGCGCGGGCGGCTCCCTGAAGTCTTGTGGTCACTTCAGTTCAGTGGTGGCTTTCCCAACGGA
1 H A E E R Y V M L P P R A R S L K S F V V T S V V A F P T D

91 TCCAACACAGTACGAGGACCCAGGACAACAGCTGCGAGTCTTGGCTGCAACCGCCGGCTGTGGAGTGTGAGCTTCAACCAACGCGGCG
31 S K A T V Q R T O D N S C F G L H A R G V E L M R F T T P G

181 TTCCCTGACAGCCCTACCCGCTCATGCCGCTGCCAGTGGGCGCTGCGGGGGACGCCAGCTCAGTGCTGAGCCTCACTTCCGACG
61 F P D S P Y P A H A R C O W A L R G D A D S V L L S L T F R S

271 TTTGACTTGGCTCTGCGACGAGCGGACGACCTGGTGACGGTGTACAACAACCTGAGCCCTACGGAGCCCAACGCGCTTGGTGAG
91 F D L A S C T D E R G S D L V T V Y N T L S P M E P H A L V O

361 TTGTGTGGCACTACCTCCCTCTTACAACCTGACCTTCCACTCTCCAGAACGCTGCTGCTATCACTGATAACCAACACTGAGCGG
121 L C G T Y P P S Y N L T F H S S O N V L L I T L I T N I E R

451 CGGCATCCCGGCTTGGAGCCACTTCTCCAGCTGCTAGGATGAGCAGCTGTGGAGGCGGCTTACGTAAAGCCGACGGGACATCAAC
151 R H P G F E A T T F O L P R M S S C G G R L R K A O G T F N

541 AGCCCTCTACTACCCAGGCACTACCCACCAACATTGACTGCACATGGAACAATGAGGTGCCAACCAACAGCATGTGAAGTGCCTTC
181 S P Y Y P G H Y P P N I D C T W N I E V P N N O H V K V R F

630 AAAITCTTCTACTGCTGGAGCGCGGCTGCTCGGGGCACTCGCCCAAGGACACTGTTGAGATCAATGGGAGAAATCTCGGGAGAG
211 K I F F Y L L E P G V P A C T C P K D Y V C I N G E K Y C G E

721 AGGTCCAGTTCGTCGTGACACGCAACAGCAACAAGATCAGAGTTCGCTTCCACTCAGATCAGTCTACACGACGACCGGCTTCTTACGT
241 R S O F V V T I S N S N K I T V R F H S D D S Y T C D T G F L A

811 GAATACCTCTCTACGACTCAGTGACCTACGCGGGGCACTTACGCTGCGCAGCGGCGGCTGTATCCGAAGGAGCTGCGCTGTGAT
271 E Y L S Y D S S D P C P G O G T T C R T G R C I R K E L R C D

901 GGCTGGGCTGACTGCACACAGCGATGACTGCACTTGGCAGCTTGGCAGCGCGCCACTGCTTACGTCGAAGAACAGTTCGCAAG
301 G W A D C T D H S D E L N C S C D A G H O F T C K N K F C K

991 CCCCTCTTCTGGGTCTGCGACAGTGTGAACGACTGCGGAGCAACAGCAGCAGCAGGGGTGCACTGTGCGGCCAGACACTTCAGGTGT
331 P L F V W V C D S V N D C G D N S D E O G C S C P A O T F R C

1081 TCCAATGGGAAGTGCCCTCTGAAAGGCGAGCAGTGAATGGGAAGGACGACTGTGGGACGGGTCCGACGAGGCTCTGCGCCCAAGGTG
361 S N G K C L L S K S O O C N G K D O C G D G S C D E A S C G C P K V

1171 AACGCTGCTACTGTACAACACACCTACCGCTCTCAATGGGCTCTGCTTGAGCAAGGGCAACCTGAGTGTGACGGGAAGGAGGAC
391 N V V T C T T G T K H T Y R C L N G L C L S K G N P E C D G K E D

1261 TGTAGCGACGGCTCAGATGAGAAGCTGCGACTTGGGCTGCGGCTATTGACGAGCAGGCTCGTGTGTGTGGGGGACGGATCGGAT
421 C S D G S D E K D C D C G L R S F T R O A R V Y G G T D A D

1351 GAGGCGAGTGCCCTGGCAGGTAAAGCTGCATCTCTGCGGCAAGGCGCACATCTGCGGTGTCTTCCCTCATCTCTCCAACTGGCTGTGCT
451 E G E W P W O V S L H A L G O G H I C G A S L I S P N W L V

1441 TCTGCCGCACTGCTACATCGATGACAGAGGATTAGGTACTCAGACCCACGAGTGGACGGCTTCTTGGGCTTGCACGACGAGAG
481 S A A H C Y I D D R G F R Y S D P T O W T A F L G L H D O S

1531 CAGCGCAGCGCCCTGGGGTGCAGGACGCGAGGCTCAAGCGCATCATCTCCACCCTTCTTCAATGACTTCACTTTCGATATGACATC
511 Q R S A P G V Q E R R L R K I I S H P F F N D A C T T F D Y D I

1621 GCGCTGTGGAGCTTGAGAAACCGGACAGATGACGCTTCCAGTGGTGGCGGCGGCTTCTGCTGCGGACGCTTCCATGTCTTCTCCGCGG
541 A L L E L E K P A E Y S S M V R P I C L P D A S H V F P A G

1711 AAGGCCATCTGGGTCTCGGGGACACACCGATGAGGAGCAGTGGCGGCTGATCTGCAAAAGGAGTGAAGTCCGCTGACATCAAC
571 K A I W V T G T G W C H T O Y G G T G A L I L O K G E I R V I N

1801 CAGACCACCTGCGAGAACCTCTGCGCAGCAGATCAGCGCGCATGATGTGCGTGGGCTTCTCAGCGCGCGGTGAGATCTTGCACG
601 Q T T C E N L L P O O I T P R M H M C V G F L S G G V G D S C D

1891 GGTGATTCGGGGGACCCCTGTCTCAGCGCTGGAGCGGATGGCGGACTCTTCCAGCGCGGTGTGGTGTGGGAGACGGCTGCGCTCAG
631 G D S G G P L S S V E A D G R I F O A G V V S W G D G C A D

1981 AGGAACAAGCAGCGGTGTACACAAGGCTCTCTGTCTTGGGACCTGGATCAAGAGAAACCTGGGGTATAGGGGCGGGGCAACCCAAA
661 R N K A C G S V Y I R L P L F R D W I K E N T G V ...

2071 TGTGTACACTGCGGGGCAACCATCTGCTCCACCTGTCAGCTGCGAGCTGGAGACTGGACGCTGACATGCAACGAGCGCCCAAC
2161 ACACATCACTGTGAACCTCAATCTCCAGGCTCCAAATGTGCTAGAAACCACTTCTGCTTCTCAGCCTCAAGTGGAGCTGGAGGTAG
2251 AAGGGGAGGACACTGGTGGTCTACTGACCAACTGGGGGCAAGGTTGAAGACACAGCTTCCCGCCGCAAGCTGGGGCGAGG
2341 CCGCTTGTGTATGTCTGCTCCCTGCTGTGAAGGACAGCGGSAAGGCTTGGAGCTCTCACTGAAGGCTGGTGGGCTGCGG
2431 ATCTGGGCTGCGGGCTTGGGCGAGCTCTTGAGCAAGGCTGCGGAGGACCTTGGGAAACAGACCGGCTTGCAGACTGAAATGAT
2521 TTACCAAGCTCCCAAGTGACTTCAGTGTGTGATGTGAATGAAGAGTAAACATTTATTTCTTTTAAAAAATAA

matriptase itself were recognized by anti-matriptase mAb 21-9 (Fig. 1B). Although sequence analysis of the 40-kDa binding protein has shown it to be a serine protease inhibitor (see below), some residual gelatinolytic activity was observed for the 95-kDa matriptase-inhibitor complex (Fig. 1C). When matriptase and its binding protein were subjected to N-terminal sequencing, only 11 amino acid residues (VVGGT-DADEGE) from matriptase were obtained, with relatively low recovery. In addition, 12 amino acid residues (GPPFAPPGL-PAG) were obtained from the amino terminus of the 40-kDa binding protein. We searched GenBank™ using these amino acid sequences for proteins related or corresponding to matriptase and its binding protein. The binding protein of matriptase was identified to be a Kunitz-type serine protease

FIG. 4. Comparison of the amino acid sequence of the C-terminal region of matriptase with trypsin, chymotrypsin, and the catalytic domains of other serine proteases. The C-terminal region (amino acids 431–683) of matriptase is compared with human trypsin (21); human chymotrypsin (22); the catalytic chains of human enteropeptidase (16), human hepsin (17), human blood coagulation factor XI (19), and human plasminogen; and the serine protease domains of two transmembrane serine proteases, human TMPRSS2 (32) and the *Drosophila* *Stubble-stubloid* gene (*Sb-sbd*) (33). Gaps to maximize homologies are indicated by *dashes*. Residues in the catalytic triads (matriptase His-484, Asp-539, and Ser-633) are boxed and indicated (▲). The conserved activation motif ((R/K)VIGG) is boxed, and the proteolytic activation site is indicated. Eight conserved cysteines needed to form four intramolecular disulfide bonds are boxed, and the likely pairings are as follows: Cys-469–Cys-485, Cys-604–Cys-618, Cys-629–Cys-658, and Cys-432–Cys-559. The disulfide bond Cys-432–Cys-559 is observed in two-chain serine proteases, but not in trypsin and chymotrypsin. Residues in the substrate pocket (Asp-627, Gly-655, and Gly-665) are boxed and indicated (♣). It is evident that the residue positioned at the bottom of the substrate pocket is Asp in trypsin-like proteases, including matriptase, but Ser in chymotrypsin.

Verification of SNC19 cDNA Encoding Matriptase—In addi-

Nucleotide and Predicted Amino Acid Sequences of a Matriptase cDNA Clone—The nucleotide and amino acid sequences of SNC19 are shown in Fig. 3. Matriptase cDNA is likely to be 2955 base pairs long when the 5'-end 33 bases and the 3'-end 92 bases from SNC19 are added to the reverse transcriptase-polymerase chain reaction fragment (2830 base pairs long). The translation initiation site was assigned to the

A LDL-receptor type regions

Matriptase (280-314)	PCPG--OFTICRTGR	CIRKELR	CDGWAD	CTDHSDELNC
(315-351)	SCDAGHOF	CKNKFCKPLFWV	CDSVNDCG	DNDSDEOGC
(352-387)	SCPA-OTFR	CSNGKCLSKSQ	QNGKDCG	GGDSDEASC
(394-430)	TCTK-HTYR	CLNGLCLSKGNPE	CDGKE	DCSDGSDEKDC
Consensus sequences				
LDL-receptor	TC---EF	C---G	CI---W	CD---DC
LRP	C---F	C---R	CI---W	CDG---DC
Perlecan	PC---P	EF	C---C	CD---D
GP-330	C---F	C---C	CI---C	CDG---DC

B C1r/s type region

Mt (1)	42	CSFGLHARGVELMRFTT	PGFDPSPYAHAR	COWALRGDADSVLSLTFRS	FDLASCDERGSDLVY
Mt (2)	168	GGRLRKAQ	GT--FNSHYYPG	HYPPNIDCTWNT	EVPPNOHVKVR
C1r (2)	193	CSSELYTEASG	Y--ISSELYPR	SYPPDLRCNYS	IRVERGLTLHLKFL
C1s (2)	175	CSGDVFTALIGE	--IASPNYFK	PYPENSRC	EYQIRLEKGFQVVTLRR
RaRF (2)	185	CSNLFRTORTGV	--ITSDFPN	PYPKSSCELYT	IELEGFMVNLQFE
CSP (2)	181	CSGDVFTALIGE	--IASPNYFK	PYPENSRC	EYQIRLEKGFQVVTLRR
Mt (1)	107	VYNTLS-PMEPHALVOL	CTYHFSYNLTFHSSQ	NVLLITLITERRH	GF 155
Mt (2)	226	PKDYVEINGEK	----YGER	--SOFVVTSSNK	ITVRFHSDQSYDTIGF 268
C1r (2)	251	PYDOLQIYANGKNIG	EFCKGRFP	DLD--TSSNAVOLL	FFIDESGDSRQW 298
C1s (2)	235	L-DSLVFVAGDRQFG	PGYCGHFG	PLNIETKSMALDI	IFOTDLTGOKKQW 283
RaRF (2)	243	PYDIKIKVGPVKVLP	GFCKEAFE	PIS--TOSHSVL	ILFHSONSGENRQW 290
CSP (2)	241	Q-DSLLFAAKNRQFG	PGFCNGFG	PLTIETHSN	ITDITVOTDLTEOKKQW 289

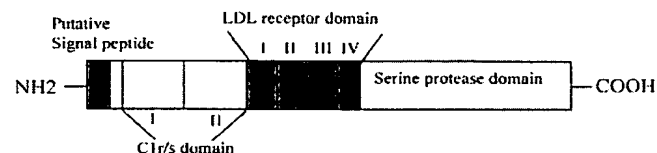


FIG. 6. Domain structure of matriptase. A schematic representation of the structure of matriptase is presented. The protease consists of 683 amino acids, and the protein product has a calculated mass of 75,626 Da. The protease contains two tandem complement subcomponent C1r and C1s domains and four tandem LDL receptor domains. The serine protease domain is at the carboxyl terminus.

fifth methionine codon because the sequence **GTCATGG** matches a favorable Kozak consensus sequence (10). This methionine is followed by four positively charged amino acids and a 14-amino acid hydrophobic region (Ser-18–Ser-31), a putative signal peptide. Assuming this methionine codon to be the initiator, the open reading frame was 2049 base pairs long, and thus, the deduced amino acid sequence was composed of 683 residues with a calculated molecular mass of 75,626 Da. The two stretches of amino acid sequences (DYVEINGEK and VVGTTDADEGE) obtained from matriptase are located in amino acids 228–236 and 443–453; thus, the translation frame is likely to be correct. There are three potential *N*-glycosylation sites with the canonical Asn-X-(Ser/Thr) sequence and an RGD sequence. An RGD sequence from proteins of the extracellular matrix has been found to mediate their interactions with integrins (11).

Structure of the Matriptase Catalytic Domain—A homology search for the deduced amino acid sequence by BLAST in the Swiss-Prot data base revealed that the carboxyl terminus at residues 432–683 of matriptase is homologous to other serine proteases and that matriptase contains the invariant catalytic triad, a characteristic disulfide bond pattern, and overall sequence similarity. Compared with the archetype serine protease chymotrypsin (12, 13) and other serine proteases, the three amino acids (His-484, Asp-539, and Ser-633) are likely to correspond to those in chymotrypsinogen (His-57, Asp-102, and Ser-195) and are likely to be essential for catalytic activity (14). The six most conserved cysteines needed to form three intramo-

lecular disulfide bonds that stabilize the catalytic pocket have been determined in other chymotrypsin-related proteases. The most likely cysteine pairings in matriptase are thus as follows: Cys-469–Cys-485, Cys-604–Cys-618, and Cys-629–Cys-658). Matriptase also contains two additional cysteines (Cys-432–Cys-559) that correspond to those used in two-chain proteases, such as enteropeptidase (15, 16), hepsin (17), plasma kallikrein (18), blood coagulation factor XI (19), and plasminogen (20), but not in trypsin (21) or chymotrypsin (22) (Fig. 4).

A putative proteolytic activation site (Arg-442) of matriptase in an Arg-Val-Val-Gly-Gly motif is similar to the characteristic RIVGG motif in other serine proteases. As mentioned above, a conserved intramolecular disulfide bond is found in those serine proteases that are synthesized as single-chain zymogens and are proteolytically activated to become active two-chain forms. This disulfide bond is proposed to hold together the active catalytic fragment with their noncatalytic N-terminal fragments. This conserved intramolecular disulfide bond has been also observed in matriptase (Cys-432–Cys-559). These sequence analyses suggest that matriptase may be synthesized as a single-chain zymogen and may become proteolytically activated to a two-chain form. If this is the case, the majority of matriptase molecules in the conditioned medium of T-47D breast cancer cells are likely to be in the zymogen form; the two-chain matriptase represents only a minor proportion of the total, consistent with the purified matriptase from T-47D human breast cancer cells exhibiting an apparent size of 80 kDa under reduced conditions (data not shown). This conclusion is also supported by the observation that the proposed N-terminal sequences for the catalytic chain of matriptase are identical to the stretch of amino acid residues (VVGTTDADEGE) that were obtained from milk-derived matriptase with very low recovery when matriptase was subjected to N-terminal sequencing.

The substrate specificity (S_1) pocket of matriptase is likely to be composed of Asp-627, positioned at its bottom, with Gly-655 and Gly-665 at its neck, indicating that matriptase is a typical trypsin-like serine protease. The predicted preferential cleavage for matriptase at amino acid residues with positively charged side chains was tested with 10 synthetic substrates with Arg and Lys residues as P1 sites. In our preliminary

studies (data not shown), matriptase was able to cleave the following synthetic substrates, presented as follows from the most rapid to the slowest: Boc-Gln-Ala-Arg-AMC, Boc-benzyl-Glu-Gly-Arg-AMC, Boc-Leu-Gly-Arg-AMC, Boc-benzyl-Asp-Pro-Arg-AMC, Boc-Phe-Ser-Arg-AMC, Boc-Val-Pro-Arg-AMC, succinyl-Ala-Phe-Lys-AMC, Boc-Leu-Arg-Arg-AMC, Boc-Gly-Lys-Arg-AMC, and Boc-Leu-Ser-Thr-Arg-AMC. Thus, matriptase may prefer substrates with amino acid residues containing small side chains, such as Ala and Gly, as P2 sites.

Structure Motifs of the Noncatalytic Region of Matriptase—The noncatalytic region of matriptase contains two sets of repeating sequences, which may serve as regulatory and/or binding domains for interactions with other proteins. Four tandem repeats of ~35 amino acids including six conserved cysteine residues (Fig. 5A) were found at the amino-terminal region (amino acids 280–430) of its serine protease domain. They are homologous to the cysteine-containing repeat of the LDL receptor (23) and related proteins (24). All of these cysteine residues are likely to be involved in disulfide bonds. In the LDL receptor, the homologous seven repeating sequences serve as the ligand-binding domain. By analogy, the four tandem cysteine-containing repeats in matriptase may also be the sites of interaction with other macromolecules. In addition, the cysteine-containing LDL receptor domain was found in other proteases such as enteropeptidase (15, 16).

The amino-terminal region of matriptase (amino acids 42–268) contains another two tandem segments with internal homology. These segments resemble partial sequences, originally identified in complement subcomponents C1r (25, 26) and C1s (27, 28). This C1r/s domain was also found in other serine proteases, such as enteropeptidase, an activator of trypsinogen (15, 16), and in the astacin subfamily of zinc metalloprotease, such as bone morphogenetic protein-1 (29) and *Drosophila tolloid* gene, a dorsal-ventral patterning protein (30). Although the exact roles of the C1r/s domains in these proteins remain unclear, a deletion of the first C1r/s domain in complement subcomponent C1r impairs tetramer formation of C1r with C1s (31). These results suggest that this domain may be involved in protein-protein interactions. In our previous study (7), a small proportion of the matriptase in breast cancer cells was identified in its complexes. One of the complexes has been isolated from human milk, and the binding protein was identified as a fragment of a Kunitz-type serine protease inhibitor. Whether the LDL receptor domain and the C1r/s domain in matriptase are both involved in the interaction with the Kunitz-type serine protease inhibitor remains to be investigated.

In conclusion, matriptase is a trypsin-like serine protease with several potential regulatory modules (Fig. 6). Its broad spectrum cleavage activity may contribute to the degradation of the extracellular matrix, activation of other proteases, and processing of growth factors. All of these ascribed functions could contribute to important aspects of tumor progression such as cancer invasion and to physiological process such as differentiation and lactation. The presence of potential protein-protein interaction domains and ligand-binding domains in matriptase suggests that the interaction of matriptase with other macromolecules on the cell surface (such as the luminal surface of the mammary gland) may regulate its activation, inhibition, and presentation. Aberrant regulation of matriptase processing may be involved in the malignant progression of cancers.

Acknowledgments—We thank Dr. Henry Yang for the automated DNA sequencing that was performed at the Lombardi Cancer Center Macromolecular Synthesis and Sequencing Shared Resource. We thank the Lombardi Cancer Center Tissue Culture Resource for cells and reagents.

REFERENCES

- Mignatti, P., and Rifkin, D. B. (1993) *Physiol. Rev.* **73**, 161–195
- Nielsen, B. S., Sehested, M., Timshel, S., Pyke, C., and Dano, K. (1996) *Lab. Invest.* **74**, 168–177
- Pyke, C., Graem, N., Ralfkiaer, E., Ronne, E., Hoyer-Hansen, G., Brunner, N., and Dano, K. (1993) *Cancer Res.* **53**, 1911–1915
- Polette, M., Gilbert, N., Stas, I., Nawrocki, B., Noel, A., Remacle, A., Stettler-Stevenson, W. G., Birembaut, P., and Foidart, M. (1994) *Virchows Arch.* **424**, 641–645
- Okada, A., Bellocq, J. P., Rouyer, N., Chenard, M. P., Rio, M. C., Chambon, P., and Basset, P. (1995) *Proc. Natl. Acad. Sci. U. S. A.* **92**, 2730–2734
- Shi, Y. E., Torri, J., Yieh, L., Wellstein, A., Lippman, M. E., and Dickson, R. B. (1993) *Cancer Res.* **53**, 1409–1415
- Lin, C.-Y., Wang, J. K., Torri, J., Dou, L., Sang, Q. A., and Dickson, R. B. (1997) *J. Biol. Chem.* **272**, 9147–9152
- Matsudaira, P. (1987) *J. Biol. Chem.* **262**, 10035–10038
- Shimomura, T., Denda, K., Kitamura, A., Kawaguchi, T., Kito, M., Kondo, J., Kagaya, S., Qin, L., Takata, H., Miyazawa, K., and Kitamura, N. (1997) *J. Biol. Chem.* **272**, 6370–6376
- Kozak, M. (1984) *Nucleic Acids Res.* **12**, 857–872
- Ruoslahti, E., and Pierschbacher, M. D. (1987) *Science* **238**, 491–497
- Hartley, B. S., and Kauffman, D. L. (1966) *Biochem. J.* **101**, 229–231
- Brown, J. R., and Hartley, B. S. (1966) *Biochem. J.* **101**, 214–228
- Hartley, B. S., Brown, J. R., Kauffman, D. L., and Smillie, L. B. (1965) *Nature* **207**, 1157–1159
- Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., and Shinomiya, K. (1994) *J. Biol. Chem.* **269**, 19976–19982
- Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W., and Sadler, J. E. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91**, 7588–7592
- Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* **27**, 1067–1074
- Chung, D. W., Fujikawa, K., McMullen, B. A., and Davie, E. W. (1986) *Biochemistry* **25**, 2410–2417
- Fujikawa, K., Chung, D. W., Hendrickson, L. E., and Davie, E. W. (1986) *Biochemistry* **25**, 2417–2424
- Forsgren, M., Raden, B., Israelsson, M., Larsson, K., and Heden, L. O. (1987) *FEBS Lett.* **213**, 254–260
- Emi, M., Nakamura, Y., Ogawa, M., Yamamoto, T., Nishide, T., Mori, T., and Matsubara, K. (1986) *Gene (Amst.)* **41**, 305–310
- Tomita, N., Izumoto, Y., Horii, A., Doi, S., Yokouchi, H., Ogawa, M., Mori, T., and Matsubara, K. (1989) *Biochem. Biophys. Res. Commun.* **158**, 569–575
- Sudhof, T. C., Goldstein, J. L., Brown, M. S., and Russell, D. W. (1985) *Science* **228**, 815–822
- Herz, J., Hamann, U., Rogne, S., Myklebost, O., Gausepohl, H., and Stanley, K. K. (1988) *EMBO J.* **7**, 4119–4127
- Leytus, S. P., Kurachi, K., Sakariassen, K. S., and Davie, E. W. (1986) *Biochemistry* **25**, 4855–4863
- Journet, A., and Tosi, M. (1986) *Biochem. J.* **240**, 783–787
- Mackinnon, C. M., Carter, P. E., Smyth, S. J., Dunbar, B., and Fothergill, J. E. (1987) *Eur. J. Biochem.* **160**, 547–553
- Tosi, M., Duponchel, C., Meo, T., and Julier, C. (1987) *Biochemistry* **26**, 8516–8524
- Wozney, J. M., Rosen, V., Celeste, A. J., Mitscock, L. M., Whitters, M. J., Kriz, R. W., Hewick, R. M., and Wang, E. A. (1988) *Science* **242**, 1528–1534
- Shimell, M. J., Ferguson, E. L., Childs, S. R., and O'Connor, M. B. (1991) *Cell* **67**, 469–481
- Cseh, S., Gal, P., Sarvari, M., Dobo, J., Lorincz, Z., Schumaker, V. N., and Zavodszky, P. (1996) *Mol. Immunol.* **33**, 351–359
- Paoloni-Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C., and Antonarakis, S. E. (1997) *Genomics* **44**, 309–320
- Appel, L. F., Prout, M., Abu-Shumays, R., Hammonds, A., Garbe, J. C., Fristrom, D., and Fristrom, J. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90**, 4937–4941
- Murdoch, A. D., Dodge, G. R., Cohen, I., Tuan, R. S., and Iozzo, R. V. (1992) *J. Biol. Chem.* **267**, 8544–8557
- Raychowdhury, R., Niles, J. L., McCluskey, R. T., and Smith, J. A. (1989) *Science* **244**, 1163–1165
- Takada, F., Takayama, Y., Hatsuse, H., and Kawakami, M. (1993) *Biochem. Biophys. Res. Commun.* **196**, 1003–1009
- Sato, T., Endo, Y., Matsushita, M., and Fujita, T. (1994) *Int. Immunol.* **6**, 665–669
- Kinoshita, H., Sakiyama, H., Tokunaga, K., Imajoh-Omhi, S., Hamada, Y., Isono, K., and Sakiyama, S. (1989) *FEBS Lett.* **250**, 411–415
- Lin, C.-Y., Anders, J., Johnson, M., and Dickson, R. B. (1999) *J. Biol. Chem.* **274**, 18237–18242

Exhibit 21

Crystal Structure of Enteropeptidase Light Chain Complexed with an Analog of the Trypsinogen Activation Peptide

Deshun Lu¹, Klaus Fütterer², Sergey Korolev², Xinglong Zheng¹
Kai Tan¹, Gabriel Waksman² and J. Evan Sadler^{1*}

¹Howard Hughes Medical
Institute, Department of
Medicine

²Department of Biochemistry
and Molecular Biophysics
Washington University School
of Medicine, 660 South Euclid
Avenue, St. Louis, MO
63110, USA

Enteropeptidase is a membrane-bound serine protease that initiates the activation of pancreatic hydrolases by cleaving and activating trypsinogen. The enzyme is remarkably specific and cleaves after lysine residues of peptidyl substrates that resemble trypsinogen activation peptides such as Val-(Asp)₄-Lys. To characterize the determinants of substrate specificity, we solved the crystal structure of the bovine enteropeptidase catalytic domain to 2.3 Å resolution in complex with the inhibitor Val-(Asp)₄-Lys-chloromethane. The catalytic mechanism and contacts with lysine at substrate position P1 are conserved with other trypsin-like serine proteases. However, the aspartyl residues at positions P2-P4 of the inhibitor interact with the enzyme surface mainly through salt bridges with the N^ε atom of Lys99. Mutation of Lys99 to Ala, or acetylation with acetic anhydride, specifically prevented the cleavage of trypsinogen or Gly-(Asp)₄-Lys-β-naphthylamide and reduced the rate of inhibition by Val-(Asp)₄-Lys-chloromethane 22 to 90-fold. For these reactions, Lys99 was calculated to account for 1.8 to 2.5 kcal mol⁻¹ of the free energy of transition state binding. Thus, a unique basic exosite on the enteropeptidase surface has evolved to facilitate the cleavage of its physiological substrate, trypsinogen.

© 1999 Academic Press

Keywords: crystal structure; enteropeptidase; serine protease; substrate recognition

*Corresponding author

Introduction

Enteropeptidase was discovered one hundred years ago in I. P. Pavlov's laboratory (Pavlov, 1902) as the first known enzyme to activate other enzymes, and it remains a remarkable example of how serine proteases have been crafted by evolution to regulate metabolic pathways. Enteropeptidase controls a primordial enzymatic cascade that is conserved among vertebrates and is essential for normal intestinal digestion. When pancreatic secretions enter the duodenum, enteropeptidase recognizes the acidic activation peptide of trypsinogen and cleaves it. The trypsin product then

cleaves and activates the other zymogens in pancreatic fluid, enabling the digestion of food. Congenital deficiency of enteropeptidase in humans causes severe intestinal malabsorption with diarrhea, vomiting, and growth failure that can be treated successfully by supplementation with pancreatic extract (Hadorn *et al.*, 1969; Haworth *et al.*, 1971).

Several enteropeptidase domains are required for the efficient activation of trypsinogen. Enteropeptidase is a two-chain polypeptide that is derived from a single-chain precursor, and consists of an N-terminal ≈120 kDa heavy chain that is disulfide-linked to a C-terminal ≈47 kDa light chain. A transmembrane segment in the heavy chain anchors enteropeptidase in the brush border of duodenal enterocytes. The light chain consists of a chymotrypsin-like serine protease domain (reviewed by Lu & Sadler, 1998). Replacement of the transmembrane domain by a cleavable signal peptide does not impair trypsinogen activation,

Present addresses: D. Lu, Cardiovascular Research Division, Eli Lilly and Company, Indianapolis, IN 46285, USA; S. Korolev, Structural Biology Center, Argonne National Laboratory, 9700 S. Cass Ave., Argonne, IL 60439, USA.

E-mail address of the corresponding author:
esadler@im.wustl.edu

indicating that membrane association is not required for substrate recognition (Lu *et al.*, 1997). The removal of heavy chain domains by reduction (Light & Fonseca, 1984), proteolysis (Mikhailova & Rumsh, 1999), or mutagenesis (LaVallie *et al.*, 1993; Lu *et al.*, 1997) reduces the rate of trypsinogen activation ≈ 500 -fold, demonstrating that the heavy chain is necessary for optimal cleavage of trypsinogen. The enteropeptidase light chain, however, is sufficient for the normal recognition of small peptidyl substrates that resemble the trypsinogen activation peptide Val-(Asp)₄-Lys (LaVallie *et al.*, 1993; Lu *et al.*, 1997).

The structural determinants of substrate specificity have not been identified on the enteropeptidase light chain, but their locations have been proposed based upon comparisons with other serine proteases. The enteropeptidase serine protease domain contains a basic tetrapeptide segment consisting of Arg96-Arg-Arg-Lys99 for porcine (Matsushima *et al.*, 1994), mouse (Yuan *et al.*, 1998), and human (Kitamoto *et al.*, 1994) enteropeptidase; or Lys96-Arg-Arg-Lys99 for bovine (Kitamoto *et al.*, 1994; LaVallie *et al.*, 1993) and rat enteropeptidase (Yahagi *et al.*, 1996). This segment is not conserved in other serine proteases, and computer modeling suggests that it is located on the protein surface where it might bind the acidic P2-P5 residues of trypsinogen activation peptides (Kitamoto *et al.*, 1994; Matsushima *et al.*, 1994) (see the legend to Figure 2 for the residue numbering). Thus, enteropeptidase appears to have an extended binding site or "exosite", distinct from the catalytic center, which recognizes substrate amino acid residues on the N-terminal side of the cleaved bond. At present there is no evidence that enteropeptidase has specificity for amino acid residues C-terminal to the scissile bond.

Similar exosites in other highly regulated serine proteases are well documented to control the recognition of substrates, cofactors and inhibitors. For example, the blood clotting protease thrombin has two so-called "anion-binding exosites" (Bode *et al.*, 1992). Exosite 1 interacts with acidic regions of preferred substrates such as fibrinogen and cofactors such as thrombomodulin. In contrast to the known properties of enteropeptidase, however, thrombin exosite 1 interacts with amino acid residues on the C-terminal side of the cleaved bond. Thrombin exosite 2 is on the opposite side of the molecule and interacts with heparin, thereby promoting the inhibition of thrombin by antithrombin (Sheehan & Sadler, 1994). These exosites have been modified by mutagenesis to create thrombin variants with novel properties (Sheehan & Sadler, 1994; Wu *et al.*, 1991). The characterization of enteropeptidase exosites, by analogous approaches, would advance our understanding of the regulation of digestion and facilitate the design of enteropeptidase derivatives with new substrate specificity.

We now have determined the crystal structure of the bovine enteropeptidase light chain complexed with an inhibitor, Val-(Asp)₄-Lys-chloromethane

(VD₄K-cm), that mimics the trypsinogen activation peptide. The catalytic mechanism and the subsite that recognizes the P1 lysine residue are conserved with other chymotrypsin-like serine proteases, but the aspartyl side-chains at positions P2-P4 of the inhibitor are accommodated mainly by ionic interactions with a unique exosite on the enzyme surface. By mutagenesis and chemical modification, we demonstrate that a single lysyl side-chain within this exosite is required for the cleavage of trypsinogen and similar peptidyl substrates. These distinctive features of enteropeptidase illustrate the specificity that serine proteases can acquire by combining modifications of the protease domain with additional motifs on accessory domains.

Results

Structure determination

The crystal structure of the serine protease domain of bovine enteropeptidase (L-BEK) bound to the inhibitor VD₄K-cm was solved by molecular replacement using the structure of γ -chymotrypsin (PDB entry code 1GCD) (Harel *et al.*, 1991) as the search model, to which enteropeptidase shows 35.9% sequence identity (Figure 1). The structure was refined to final *R* factors of *R* = 23.4% and *R*_{free} = 26.9% (Figure 2 and Table 1). For ease of comparison to related serine protease structures, we use the chymotrypsin-derived residue numbering scheme proposed by Bode *et al.* (1992). The protein used for the present structure determination (L-BEK) contains only 13 C-terminal amino acid residues of the enteropeptidase heavy chain. Note that the usage of the terms "heavy" and "light" chain is the reverse of what is common usage for chymotrypsin and thrombin. The present structure shows an uninterrupted backbone for the two-chain molecule, comprising residues 1 through 7 (chymotrypsin numbering) of the N-terminal domain and residues 16 through 243 of the serine protease domain. Residues 8 through 13 of the N-terminal domain and residues 244 and 245 of the serine protease domain protrude freely into the solvent and could not be modeled.

Tertiary structure

As expected, based upon its homology to other serine proteases, L-BEK is very similar in fold to both representative family members chymotrypsin and thrombin (Figure 3(a) and (c)): the tertiary structure consists of two six-stranded β -barrels, either of which makes up about one half of the entire molecule. The structure of L-BEK superimposes on chymotrypsin with a root-mean-square deviation of 1.10 Å for 224 C α positions, and it superimposes on thrombin with a root-mean-square deviation of 1.23 Å for 234 C α positions. Variations in secondary structure occur mainly in the loop regions. L-BEK also contains, relative to

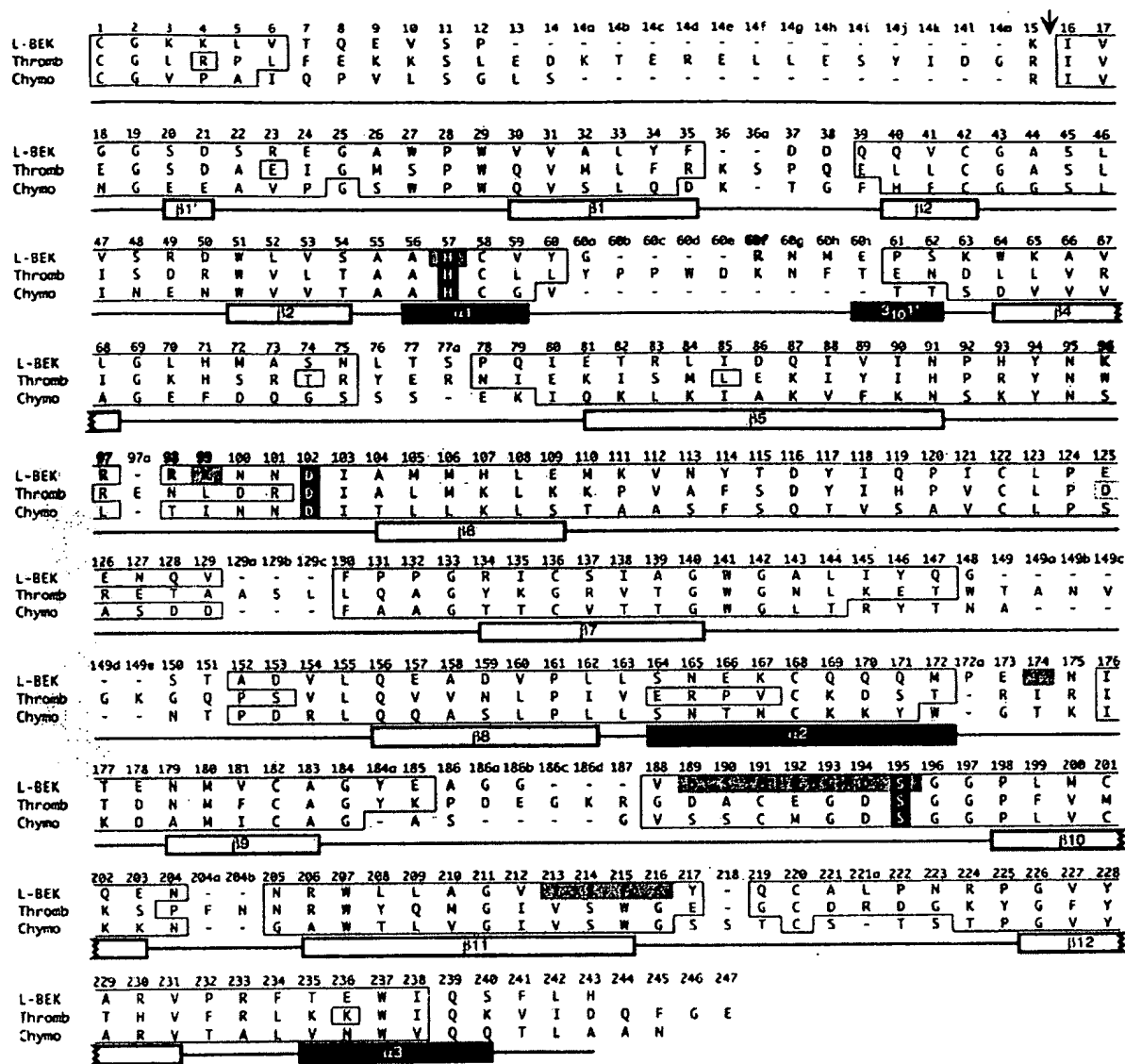


Figure 1. Sequence alignment of enteropeptidase (L-BEK), chymotrypsin (Chymo) and thrombin (Thromb) protease domains. Amino acid sequences are aligned based on topological equivalence of the superimposed crystal structures. Amino acid residues are numbered based on the sequence of chymotrypsinogen. Residues of L-BEK and the other proteases are boxed if the separation between C α positions is ≤ 1.6 Å. Active-site residues (His57, Asp102, Ser195) are in filled black boxes. Residues in contact with the VD₄K-cm inhibitor are shaded in blue. L-BEK secondary structure elements are indicated below the sequences; helices (α -helix, 3_{10} -helix) are shown as filled boxes and β -strands are shown as open boxes. Secondary structure conserved with γ -chymotrypsin are numbered sequentially, and those designated by prime numbers (i.e. $3_{10}1'$, $\beta 1'$) are not present in γ -chymotrypsin. The arrow indicates the activation cleavage site that separates the heavy chain remnant (residues 1-15) from the light chain (residues 16-243).

chymotrypsin, an additional β -strand, $\beta 1'$, and an additional small 3_{10} -helix, $3_{10}1'$ (Figures 1 and 3(a)). The 3_{10} -helix is part of the so-called "60-loop" that connects helix $\alpha 1$ and strand $\beta 4$, and a similar 3_{10} -helix is present in the much longer 60-loop of thrombin.

The enteropeptidase serine protease domain is stabilized by five disulfide bonds, all of which are

conserved with chymotrypsin: Cys1-Cys122, Cys42-Cys58, Cys136-Cys201, Cys168-Cys182, and Cys191-Cys220 (Figure 3(a)). Thrombin lacks one of these disulfide bonds, corresponding to that between Cys136 and Cys201 of enteropeptidase. The 13 residue N-terminal chain of L-BEK is covalently linked to the serine protease domain by the disulfide bond between Cys1 and Cys122.

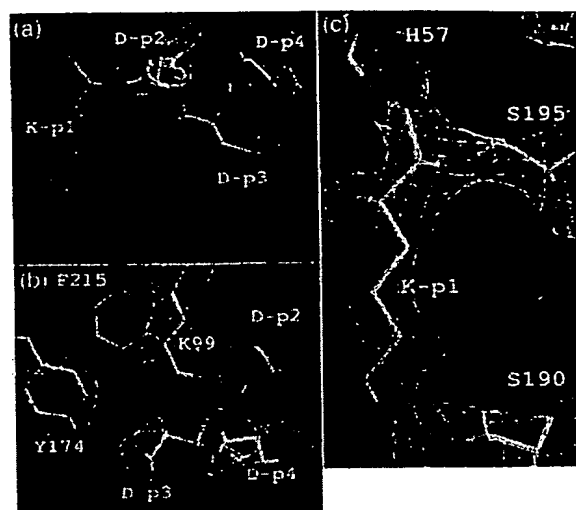


Figure 2. Representative regions of electron density. Simulated annealing omit maps, using Fourier coefficients $F_o - F_c$ and model phases, were calculated by deleting the VD₄K-chloromethane inhibitor either (a) alone or (b)-(c) including an additional region of 3.5 Å around it. (a) View of the inhibitor peptide from the protein outwards. Electron density for the hexapeptide is observed for positions P1 to P4. (Amino acid residues of peptidyl substrates or inhibitors customarily are numbered P1, P2, P3, etc., from the scissile bond toward the N terminus, and P1', P2', on the C-terminal side of the scissile bond. The corresponding subsites on the cognate protease are numbered S1, S2, S3 and S1', S2' (Schechter & Berger, 1967)). (b) Interaction of the aspartyl side-chains of residues P2-P4 with Lys99 and Tyr174 of L-BEK. (c) Covalent linkage of the C terminus of the inhibitor to the catalytic residues His57 (N^{ε2}-methylene carbon) and Ser195 (O^γ carbonyl carbon atom), mimicking the tetrahedral intermediate of the hydrolysis reaction. The figure was produced with the program O (Jones & Thirup, 1986; Jones *et al.*, 1991).

Aside from this single disulfide bond, the interactions of this short polypeptide with the bulk of the structure are relatively weak, consisting of an amino-aromatic interaction between Lys4 and Trp27, and hydrogen bonds between main-chain atoms of Gly2 and either Trp207 or Pro120. Consequently, the remaining residues 8-13 of the heavy chain are disordered.

The catalytic center

The catalytic center contains the signature structural elements of serine proteases: the catalytic triad consisting of Asp102, His57 and Ser195; the oxyanion hole formed by the main-chain amide nitrogen atoms of residues 193 and 195; and the S1 subsite or specificity pocket that interacts with the side-chain of the P1 substrate/inhibitor residue (Figure 4(a) and (d)). The VD₄K-cm inhibitor is

Table 1. Data collection and refinement statistics

A. Data collection	
Data set	Native
Radiation, detector system	CuKα, Raxis
Resolution (Å)	30-2.3
Total/unique reflections	28,051/10,541
Completeness (%) ^a	92.6 (89.2)
R _{sym} (%) ^b	4.4 (8.8)
B. Refinement	
Resolution (Å)	30.0-2.3
Reflections (completeness) ^c (%)	9854 (87.6/82.0)
Non-H atoms	2023
R/R _{free} (%) ^d	23.4/26.9
r.m.s. deviations ^e	
Bond lengths (Å)	0.006
Bond angles (deg.)	1.39
B values (main-chain/side-chain) (Å ²)	1.5/2.0

^a Completeness for $I/\sigma(I) > 1.0$; value for high resolution shell (2.38-2.3 Å) in parentheses.

^b $R_{sym} = \sum |I - \langle I \rangle| / \sum I$, where I = observed intensity, and $\langle I \rangle$ = average intensity from multiple observations of symmetry-related reflections; the value for the high-resolution shell is in parentheses.

^c Numbers reflect the "working set" of reflections at $F/\sigma(F) > 2.0$; values for completeness for the overall/high-resolution shell (2.4-2.3 Å) are in parentheses.

^d R_{free} was calculated on the basis of 546 reflections (5.5% of the observed reflections) that were randomly omitted from the refinement.

^e Root-mean-square (r.m.s.) deviation from ideal bond lengths and angles (Engh & Huber, 1991) and r.m.s. deviation in B-factors of bonded atoms.

identical in sequence to the trypsinogen activation peptide and is covalently bound to the catalytic residues His57 and Ser195 through its C-terminal residue Lys-P1 (Figures 2(c) and 4(a)). The carbonyl carbon atom of Lys-P1 forms a tetrahedral hemiketal with Ser195 O^γ, and the methylene carbon atom of the inhibitor is bound to the imidazole ring (N^{ε2}) of His57. This arrangement mimics the tetrahedral intermediate of the substrate hydrolysis reaction. The side-chain of Lys-P1 inserts deeply into the S1 pocket, at the bottom of which Asp189 neutralizes the terminal amino group (Figure 4(b)). The interactions of Lys-P1 at the bottom of the specificity pocket also include short hydrogen bonds to both the hydroxyl group and the carbonyl oxygen atom of Ser190. Lys-P1 also makes short hydrogen bonds to two water molecules, WAT438 and WAT407, that correspond to water molecules 429 and 494, respectively, of the thrombin-hirugen complex (Vijayalakshmi *et al.*, 1994). These two water molecules are conserved among several serine protease structures (Krem & Di Cera, 1998). The aliphatic part of the Lys-P1 side-chain packs against the main-chain atoms of Phe215 and Ser214, as well as the C^{γ2} atom of Thr213 (Figure 4(b) and (d)).

The extended substrate binding exosite

Despite its covalent attachment to the protein through the catalytic center, the VD₄K-cm inhibitor is disordered at its N-terminal end and electron

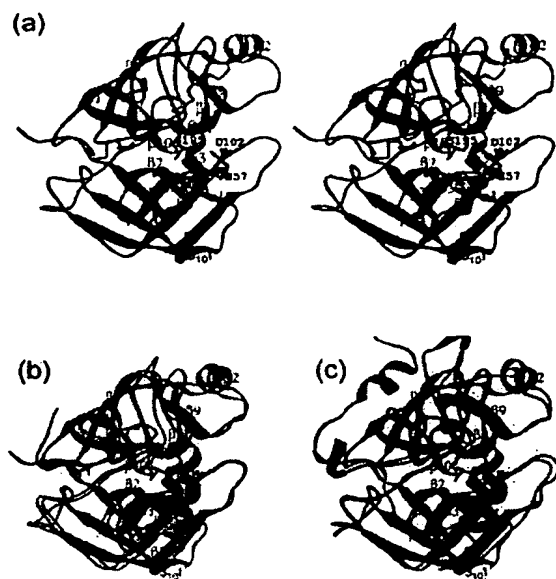


Figure 3. Overall fold of enteropeptidase compared to γ -chymotrypsin and α -thrombin. (a) Stereo ribbon diagram of L-BEK. The catalytic residues are labeled and the disulfide bonds are shown in yellow. Superposition of L-BEK (grey) with (b) γ -chymotrypsin (1GCD, in cyan) and (c) with human α -thrombin (1PPB, in green). The structures were aligned with respect to the C^α positions of the catalytic residues His57, Asp102 and Ser195, and are shown in the same orientation as for L-BEK in (a). This Figure was produced with the program RIBBONS (Carson, 1997), as were Figures 4(a)(c), 5, and 7.

density was observed only for residues Lys-P1 through Asp-P4 (Figure 2(a)). The inhibitor geometry is remarkably similar to that of D-Phe-Pro-Arg-chloromethane (PPACK) in thrombin, as illustrated in Figure 5. The alignment of L-BEK with thrombin, based only on the C^α atoms of the catalytic triad, leads to a near perfect superposition of the two inhibitor molecules, including the C^β positions, despite their complete lack of sequence similarity. Although VD₄K-cm forms two main-chain to main-chain hydrogen bonds with residues in strand β 11 (Figures 1 and 4(d)), it does not otherwise adopt a β -strand configuration in contrast to what is observed for the thrombin-PPACK structure (Bode *et al.*, 1992).

Aside from the S1 subsite, the major determinant of VD₄K-cm recognition is Lys99. The basic side-chain of this residue coordinates the aspartic acid side-chains at positions P2 through P4 of the inhibitor. These three carboxylate groups surround the terminal amino-group of Lys99 in a fashion similar to an inverted tripod. Lys99 forms salt bridges only with Asp-P2 and Asp-P4, whereas Asp-P3 is hydrogen bonded to the hydroxyl

moiety of Tyr174 (Figure 4(c) and (d)). Residue Phe215 is also indirectly involved in substrate binding, with its phenyl ring serving as a hydrophobic platform that supports the side-chain of Lys99 (Figures 2(b) and 4(c)).

Lys99 is part of a sequence of four basic amino acid residues in the β 5 β 6 loop that, based on molecular modeling, had been predicted to define the substrate specificity of enteropeptidase (Kitamoto *et al.*, 1994; Matsushima *et al.*, 1994). In the present crystal structure the side-chain of Arg97 is completely disordered, that of Arg98 is poorly defined, and both extend into solvent. Lys96 does not make any close contacts with the inhibitor, but folds back onto the protein surface to form a short hydrogen bond (2.8 Å) with the hydroxyl group of Tyr94. Tyr60 also is in close proximity to the terminal amino group of Lys96. As discussed below, the contribution of these basic residues to substrate recognition was examined further by mutagenesis.

The electrostatic surface of L-BEK (Figure 6) includes two prominent positive charges in the vicinity of the inhibitor binding site: Lys99 is on the N-terminal side and Arg60f is on the C-terminal side of the scissile bond position. Arg60f is held in place by hydrophobic interactions with the aromatic ring of Phe35 and a short hydrogen bond donated by the carbonyl oxygen atom of Cys58 (Figure 7). The latter interaction positions the guanidinium group of Arg60f at a distance of 8 Å from the catalytic center, where it would not be expected to have a direct effect on the recognition of VD₄K-cm. In the superposition with thrombin (Figure 7), the C^α atom of Arg60f is closest to the C^α atom of Phe60h, but its guanidinium group lies close to the head group of Lys60f; the latter forms a hydrogen bond with the carbonyl oxygen atom of His57. The basic nature of these side-chains and their similar position relative to the catalytic center suggest that Arg60f of enteropeptidase and Lys60f of thrombin may have a similar function in recognition of residues C-terminal to the scissile bond. For thrombin, the effects of mutagenesis are consistent with this hypothesis because alteration of Lys60f markedly impairs the cleavage of fibrinogen without affecting the cleavage of D-Phe-pipecolyl-Arg-p-nitroanilide (Wu *et al.*, 1991).

Mutagenesis and chemical modification of L-BEK

To determine the contribution of specific basic amino acid residues to substrate recognition, mutant forms of L-BEK were prepared in which each of the Arg or Lys residues at positions 60f and 96-99 was changed to Ala. The proteins were expressed in a baculovirus system and purified by affinity chromatography on STI-agarose. In addition, a sample of purified L-BEK was treated with acetic anhydride. The conditions of acetylation were shown previously to result in the efficient modification of lysyl residues on porcine enteropeptidase (Baratti & Maroux, 1976). By

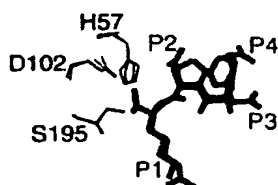


Figure 5. Structural superposition of the VD_4K -cm inhibitor of enteropeptidase with D-Phe-Pro-Arg-chloromethane (PPACK) of thrombin (1PPB). The alignment resulted from the superposition of the C^α positions of the catalytic residues His57, Asp102, and Ser195 in both proteins. Enteropeptidase residues and inhibitor atoms are shown in color-coded sticks: grey for C, red for O, blue for N. Residues and inhibitor atoms of thrombin are shown in green sticks. The view is from the protein outwards.

side-chain increases ΔG_T by 2.1 to 2.5 kcal mol⁻¹. Acetylation of L-BEK also markedly decreased the rate of cleavage of both GD_4K -na ($\approx 1.5\%$) and trypsinogen ($\approx 1.5\%$), but enhanced the cleavage of Z-Lys-SBzl (Figure 9 and Table 2).

Rate constants for inhibition by VD_4K -cm also were determined to assess the effect of mutations on the recognition of the trypsinogen activation peptide (Table 3). The magnitude and direction of the changes are similar to those observed for cleavage of GD_4K -na and trypsinogen. The substitutions Arg60fAla, Lys96Ala, Arg97Ala, and Arg98Ala had modest effects on the inhibition reaction, increasing ΔG_T by 0.3 to 0.8 kcal mol⁻¹. In contrast, the mutation Lys99Ala markedly reduced the rate of inhibition, increasing ΔG_T by 1.8 kcal mol⁻¹. Acetylation of L-BEK also markedly slowed the rate of inhibition by VD_4K -cm, increasing ΔG_T by 2.7 kcal mol⁻¹. These values of $\Delta\Delta G_T$ for inhibition by VD_4K -cm are consistent with

those estimated from the relative rates of substrate cleavage (Figure 9).

Discussion

Structural Interpretation of substrate specificity

Limited qualitative studies employing protein substrates (Anderson *et al.*, 1977; Light *et al.*, 1980) and synthetic peptides (Maroux *et al.*, 1971) indicate that mammalian enteropeptidase is remarkably specific. With few exceptions, the P1 residue must be basic (e.g. Lys, Arg, or homoarginine) and the P2 and P3 positions must be acidic (e.g. Asp, Glu or carboxymethylcysteine). The substituents at P4 and P5 are less critical, but additional acidic residues in these positions increase affinity for the enzyme (Maroux *et al.*, 1971).

The crystal structure of L-BEK provides a reasonable explanation for these properties. The catalytic center of enteropeptidase is conserved with related enzymes that prefer a basic side-chain in the P1 position such as trypsin, and Lys-P1 of the inhibitor VD_4K -cm makes numerous close contacts with L-BEK (Figure 4(d)). Acidic residues on the N-terminal side of residue P1 interact with an extended exosite on the enzyme surface, and the number of contacts decreases as the distance from the catalytic center increases. For example, Asp-P2 main-chain atoms make four close contacts with L-BEK, and its carboxylate side-chain makes two H-bonds with the N^ϵ atom of Lys99; Asp-P3 makes half as many contacts, Asp-P4 makes only one H-bond between its carboxylate group and the N^ϵ atom of Lys99, and residues Asp-P5 and Val-P6 are disordered. Thus, the interface between L-BEK and VD_4K -cm is consistent with the increased tolerance for variations in substrate structure at positions distal to P3.

The distribution of interactions between VD_4K -cm and bovine enteropeptidase is mirrored by the observed variation among trypsinogen activation peptides. Sequences are known for at least 30



Figure 6. Electrostatic surface diagram of the Val-(Asp)₄-Lys-chloromethane inhibitor binding site of enteropeptidase. Negative and positive surface charges are shown in deep red and blue, respectively, with linear interpolation in between. Conserved water molecules WAT407 and WAT438 are shown as spheres in cyan, inhibitor atoms are shown as sticks and are color-coded as described in the legend to Figure 4. (a) Overall view. (b) Close up view of the S1 binding pocket. The Figure was produced with the program GRASP (Nicholls *et al.*, 1991).

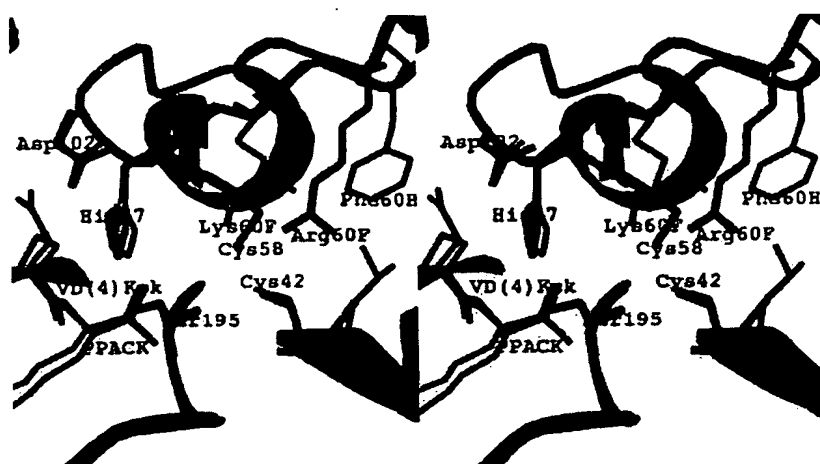


Figure 7. Structural role of residue Arg60f in comparison to Lys60f of thrombin. Enteropeptidase secondary structure elements are shown in grey and atoms are color-coded as described in the legend to Figure 4. Secondary structure elements and carbon atoms of thrombin are shown in green, keeping all other atom color assignments unaltered. The structures were aligned as shown in Figure 3. Interestingly, Arg60f aligns with Phe60h of thrombin with regard to the C α position, while its guanidinium group is very close to the terminal amino group of Lys60f of thrombin.

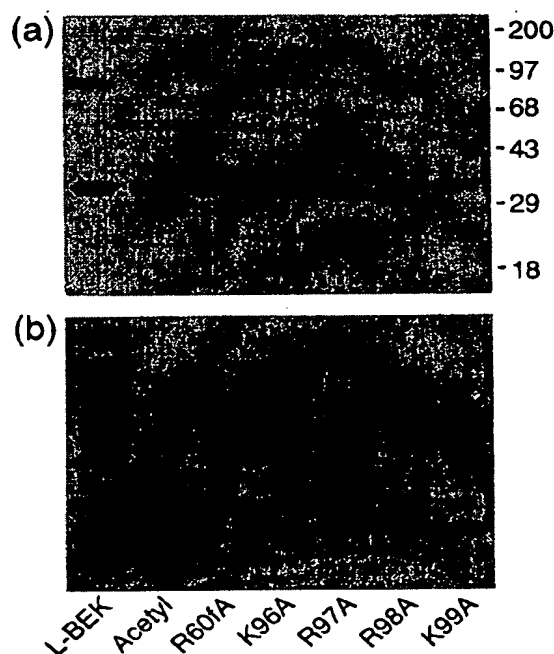


Figure 8. Gel electrophoresis of enteropeptidase variants. (a) Samples (5 μ g) of affinity purified enteropeptidase variants were analyzed by SDS-polyacrylamide gel electrophoresis without reducing agent and visualized by staining with Coomassie brilliant blue (Laemmli, 1970). The positions of molecular mass markers are indicated at the right in kilodaltons. (b) Enteropeptidase variants were analyzed by native gel electrophoresis using a similar polyacrylamide gel and buffer system except that SDS was omitted from the sample buffer.

genetically distinct trypsinogens, representing mammals, birds, amphibians and fish (Bricteux-Gregoire *et al.*, 1972; Lu & Sadler, 1998). Position P1 is occupied almost exclusively by Lys. Very few trypsinogens have Glu instead of Asp at position P2 or P3. Most residues at position P4 are Asp, but Glu or Asn occur in $\approx 30\%$ of cases. Position P5 shows more variation; Asp is present in $\approx 60\%$, but aromatic, aliphatic, small polar and basic side-chains also are found. Position P6 is not conserved. Therefore, the tendency of trypsinogen activation peptide residues to vary during vertebrate evolution correlates inversely with the number and location of close contacts in the L-BEK-VD $_4$ K structure.

Energetic contributions of specific residues to substrate recognition

The contacts between L-BEK and VD $_4$ K-cm are dominated by ionic interactions between aspartyl side-chains and Lys99, and the importance of these interactions is supported by the effect of acetylation on enteropeptidase specificity. Reaction of porcine enteropeptidase with acetic anhydride reduces its activity toward trypsinogen by more than 98%, but increases its activity toward L-N- α -benzoylarginine *p*-nitroanilide (L-BAPNA) by 1.8-fold (Baratti & Maroux, 1976). These studies were performed with full-length enteropeptidase and therefore could not localize the critical modified residues to either the light chain or the heavy chain. However, we found that acetylated L-BEK has a similar phenotype: it cleaves the simple thioester substrate Z-Lys-SBzl more rapidly than does native L-BEK (Table 2), but cannot cleave either GD $_4$ K-na or trypsinogen (Figure 9). Thus,

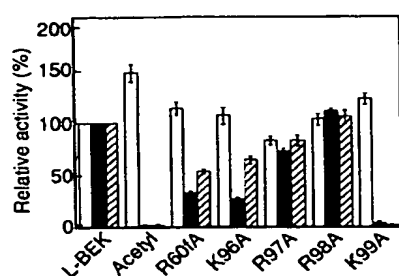


Figure 9. Relative rates of substrate cleavage by enteropeptidase variants. The activity of the indicated preparations of enteropeptidase light chain was assayed with the substrates Z-Lys-SBzl (open boxes), GD₄K-na (filled boxes), and trypsinogen (hatched boxes). The values obtained are expressed as the mean percentage \pm SE for at least three independent determinations, normalized to the activity observed for wild-type L-BEK (100 %).

residues in the enteropeptidase light chain that are sensitive to acetylation, such as Lys or Tyr, are necessary for the recognition of peptidyl substrates. The best candidate target to explain the effect of acetylation is Lys99, which makes at least three H-bonds with Asp-P2 and Asp-P4 in the L-BEK-VD₄K complex (Figure 4(d)). The other possibility, Tyr174, makes only a single H-bond with Asp-P3.

Mutagenesis and kinetic studies support a major contribution of Lys99 to the energetics of substrate binding. Substitution of Lys99 by alanine caused similar impairments in the ability of enteropeptidase to cleave either GD₄K-na or trypsinogen (Figure 9 and Table 2), and in the rate of enteropeptidase inhibition by VD₄K-cm (Table 3). For the latter reaction, the Lys99Ala mutation increased ΔG_T by 1.8 kcal mol⁻¹ and acetylation of L-BEK increased ΔG_T by 2.7 kcal mol⁻¹. Mutations at other positively charged residues have much smaller effects on the kinetics of substrate cleavage or inhibition by VD₄K-cm. The similar phenotypes of acetylated L-BEK and the Lys99Ala mutant are consistent with the importance of ionic interactions in the recognition of substrate residues in the P2-P4 positions, and suggest that the effects of

acetylation are due mainly to the loss of positive charge at Lys99.

A hierarchy of functional sites participates in substrate recognition

The extended contacts between L-BEK and VD₄K-cm appear to explain the preference of enteropeptidase for similar peptidyl substrates, but do not fully account for the efficient activation of trypsinogen. Two-chain enteropeptidase cleaves trypsinogen \approx 500-fold more rapidly than does the isolated light chain (Lu *et al.*, 1997), indicating that the heavy chain promotes physiological substrate recognition. Thus, a hierarchy of functional sites has evolved to optimize trypsinogen activation. The catalytic center confers specificity for cleavage after basic amino acid residues. An exosite on the light chain, distinct from the catalytic center, recognizes acidic trypsinogen activation peptides, and at least one site on the heavy chain interacts with and further accelerates the cleavage of trypsinogen. This feature of the enteropeptidase-trypsinogen interaction is shared by many other serine proteases that participate in highly regulated metabolic pathways, and it illustrates general principles underlying the adaptation of serine proteases to cleave a restricted range of substrates. Such adaptation often has been accomplished by exploiting structural features of both catalytic and non-catalytic domains to interact with complementary surfaces on cofactors or substrates.

Materials and Methods

Reagents and proteins

Bovine trypsinogen and bovine trypsin were from Worthington (Freehold, NJ). Thiobenzyl benzyloxycarbonyl-L-lysinate (Z-Lys-SBzl), and the enteropeptidase substrate Gly-Asp-Asp-Asp-Asp-Lys- β -naphthylamide (GD₄K-na) were from Bachem (King of Prussia, PA). Chromogenic substrates S-2366 (pyroGlu-Pro-Arg-p-nitroanilide) and S-2765 (Z-D-Arg-Gly-Arg-p-nitroanilide) were from Chromogenix (Sweden). Ovomucoid, soybean trypsin inhibitor agarose (STI-agarose), acetic anhydride, p-nitrophenyl p'-guanidinobenzoate, and 5,5'-dithiobis(2-nitrobenzoic acid) (DTNB) were from Sigma (St. Louis, MO).

Table 2. Kinetic parameters for the cleavage of substrates Z-Lys-SBzl and GD₄K-na

Enzyme	Z-Lys-SBzl			GD ₄ K-na		
	K_m (μ M)	k_{cat} (s ⁻¹)	k_{cat}/K_m (μ M ⁻¹ s ⁻¹)	K_m (mM)	k_{cat} (s ⁻¹)	k_{cat}/K_m (mM ⁻¹ s ⁻¹)
L-BEK	120 \pm 10	129 \pm 4	1.05	0.61 \pm 0.09	42.7 \pm 4.0	70.4
Acetyl L-BEK	40 \pm 10	111 \pm 4	2.93	NA	NA	NA
R60fA	120 \pm 10	159 \pm 19	1.36	0.73 \pm 0.08	12.7 \pm 1.0	17.3
K96A	100 \pm 30	108 \pm 22	1.10	1.25 \pm 0.07	17.1 \pm 1.5	13.7
R97A	120 \pm 40	128 \pm 33	1.02	0.66 \pm 0.07	25.5 \pm 2.3	38.6
R98A	140 \pm 10	128 \pm 3	0.88	0.77 \pm 0.02	39.1 \pm 0.8	51.0
K99A	50 \pm 10	120 \pm 1	2.53	NA	NA	NA

Values for K_m and k_{cat} are expressed as the mean \pm SE of three independent determinations. NA, activity insufficient to determine kinetic constants.

Table 3. Kinetic parameters for the inhibition of enteropeptidase by VD₄K-cm

Enzyme	k_2 (s ⁻¹)	K_i (μM)	k_2/K_i (mM ⁻¹ s ⁻¹)	$\Delta\Delta G_T$ (kcal mol ⁻¹)
L-BEK	0.013 ± 0.003	1.0 ± 0.3	13.4 ± 2.5	0
Acetyl L-BEK	0.0010 ± 0.0001	7.3 ± 1.2	0.15 ± 0.02	+2.7
R60fA	0.061 ± 0.015	17 ± 5	3.59 ± 0.08	+0.8
K96A	0.0048 ± 0.0008	0.9 ± 0.3	5.9 ± 1.3	+0.5
R97A	0.0073 ± 0.0015	1.0 ± 0.3	7.5 ± 0.4	+0.3
R98A	0.0072 ± 0.0002	0.84 ± 0.04	8.7 ± 0.2	+0.3
K99A	0.00024 ± 0.00001	0.4 ± 0.2	0.6 ± 0.1	+1.8

Values for K_i and k_2 are expressed as the mean ± SE of at least three independent determinations.

Plasmid constructs

Plasmid pBlue-newL was prepared from pBEK by a PCR mutagenesis strategy as described (Lu *et al.*, 1997; Nelson & Long, 1989) and encodes the human prothrombin signal peptide (Met1-Phe28) fused to the carboxyl-terminal 251 amino acid residues of bovine enteropeptidase (Tyr785-His1035) (Kitamoto *et al.*, 1994). Using a similar mutagenesis method, plasmid pBlue-newL was altered to contain mutations encoding each of the amino acid substitutions Arg60fAla, Lys96Ala, Arg97Ala, Arg98Ala, and Lys99Ala. The segment encoding the chimeric prothrombin-enteropeptidase construct was excised from each plasmid by digestion with *Hind*III, made blunt with DNA polymerase, and ligated into the *Sma*I site of the expression vector pVL1392 (Pharmingen, Carlen, CA) to yield plasmids pVLnewL, pVLR60fA, pVLK96A, pVLR97A, pVLR98A, and pVLK99A.

A fragment of plasmid pBEK encoding amino acid residues Cys788-His1035 of bovine enteropeptidase (Kitamoto *et al.*, 1994) was amplified by PCR and inserted into the *Nco*I site of expression vector pET-11d (Novagen, Madison, WI) to yield plasmid pETL. The construct encodes two amino acid residues derived from the vector (Met-Ala) before commencing with enteropeptidase sequence at Cys788. For all plasmids, the segments derived by PCR were sequenced to confirm the accuracy of the construction.

Production of enteropeptidase light chain in *Escherichia coli* (L-BEK)

E. coli BL21 (DE3) cells (Stratagene) containing pETL were grown in two liters of LB/ampicillin medium, and recombinant L-BEK was solubilized from the inclusion bodies at room temperature with 10 ml of 0.1 M Tris-HCl (pH 8.6), 1 mM EDTA-Na, 150 mM dithioerythritol, and 6 M guanidine HCl. L-BEK was refolded by a modification of a protocol described for the refolding of tissue plasminogen activator from lysates of *E. coli* (Kohnert *et al.*, 1992). After centrifugation for 30 minutes at 50,000 g, the solubilized protein was dialyzed at room temperature against 3 M guanidine-HCl (pH 2.5), and mixed with 10 ml of oxidation buffer (50 mM Tris-HCl (pH 9.3), 6 M guanidine-HCl, 0.1 M oxidized glutathione). After dialysis against 3 M guanidine-HCl (pH 8.0), disulfide exchange and refolding were initiated by dropwise dilution with stirring into 500 ml of 0.7 M arginine-HCl (pH 8.6), 2 mM reduced glutathione, and 1 mM EDTA. After 72 hours, the reaction was dialyzed against 20 mM Tris-HCl (pH 7.6), 20 mM NaCl, and then digested with trypsin (1:50 molar ratio) for one hour. The trypsin was inactivated with a fourfold excess of ovomucoid and active L-BEK was purified to homogeneity by affinity

chromatography on STI-agarose. The yield was 10 mg per two liter culture.

The N-terminal amino acid sequence of L-BEK was determined after SDS-PAGE and electroblotting onto a polyvinylidene difluoride membrane (Kalafatis & Mann, 1993). The product had the expected two-chain structure and the predicted first Met residue was removed completely during biosynthesis. The mass of L-BEK was 27,741 Da by electrospray ionization mass spectrometry, and this value is consistent with the calculated mass of 27,739.6 Da. The concentration of L-BEK determined by active-site titration with *p*-nitrophenyl *p*'-guanidinobenzoate (Chase & Shaw, 1970) agreed with the value determined spectrophotometrically at 280 nm using the calculated extinction coefficient (Pace *et al.*, 1995) of 70,870 M⁻¹cm⁻¹.

Production of wild-type and mutant enteropeptidase in baculovirus

Constructs pVLnewL, pVLR60fA, pVLK96A, pVLR97A, pVLR98A, and pVLK99A were cotransfected with BaculoGold DNA (Pharmingen) into Sf9 cells and high-titer recombinant baculovirus was prepared by repeated infection. High Five cells (1 × 10⁶ per ml, Invitrogen) were grown in Express Five serum free medium supplemented with 20 mM glutamine. Suspension cultures (200 ml each) were infected with 0.5 ml virus stock. After 72 hours, conditioned medium was collected and adjusted to pH 8.0 by addition of ≈20 ml/l 1 M Tris-HCl (pH 8), and precipitated glutamine was removed by centrifugation. Recombinant enteropeptidase was purified by affinity chromatography on STI-agarose. The yield was up to ≈15 mg of apparently homogeneous enteropeptidase light chain per liter of medium.

Affinity purification of enteropeptidase light chain variants on STI-agarose

High Five cell conditioned medium (1000 ml) was applied at 50 ml/hour to a column (2 ml) of STI-agarose equilibrated with 20 mM Tris-HCl (pH 7.5), 50 mM NaCl, at 4°C. The column was washed with 10 ml of 20 mM Tris-HCl (pH 7.5), 1 M NaCl, followed by 50 ml of 20 mM Tris-HCl (pH 7.5). Enteropeptidase was eluted with 50 mM glycine-HCl (pH 3.0); 1 ml fractions were collected and neutralized immediately with 50 μl of 2 M Tris-HCl (pH 8.0). Refolded and trypsin-activated L-BEK prepared in *E. coli* was purified similarly, applying the product obtained from a two liter culture to the column. Fractions were analyzed by SDS-PAGE (Laemmli, 1970) and silver staining (Morrissey, 1981), pooled, dialyzed

against 20 mM Tris-HCl (pH 7.5), 50 mM NaCl, and stored at -70°C .

Preparation of a stoichiometric complex of L-BEK and VDDDDK-chloromethane

The active site directed inhibitor Val-(Asp)₄-Lys-chloromethane (VD₄K-cm) was synthesized (Haematologic Technologies, Inc.) and its structure was confirmed by amino acid composition. Electrospray ionization mass spectrometry gave a mass of 739.3 Da and the predicted mass was 739.2 Da. Affinity-purified L-BEK from *E. coli* (10 mg) in 100 ml of 20 mM Tris-HCl (pH 7.5), 50 mM NaCl, was reacted on ice with 50 ml of 100 μM VD₄K-cm added dropwise over 60 minutes. The L-BEK-VD₄K complex was dialyzed at 4°C against 20 mM Tris-HCl (pH 7.5), 50 mM NaCl, and concentrated to 25 mg/ml by ultrafiltration (Centricon-30, Amicon). The mass determined by electrospray ionization mass spectrometry (28,448 Da) was consistent with the mass calculated for the expected stoichiometric complex (28,442.3 Da).

Crystallization of L-BEK and data collection

Crystals of L-BEK-VD₄K complex were grown at 20°C in a hanging drop against a reservoir of 100 mM sodium cacodylate (pH 5.0), 10 mM zinc sulfate, and 10% (w/v) PEG-400 at a protein concentration of 4 mg/ml. The crystals were orthorhombic ($P2_12_12_1$) with one molecule per asymmetric unit and cell dimensions of $a = 39.99 \text{ \AA}$, $b = 70.65 \text{ \AA}$, and $c = 85.22 \text{ \AA}$. A crystal was transferred into cryoprotectant buffer containing 100 mM sodium cacodylate (pH 5.0), 20 mM zinc sulfate and 25% (w/v) PEG-400, and frozen at 100 K in a stream of nitrogen vapor. Data were collected using a Rigaku RaxisII image plate detector mounted on a Rigaku RU200 rotating copper anode. A data set complete to 2.3 \AA resolution was collected. Data were processed and scaled using the programs DENZO and SCALEPACK (Otwinowski & Minor, 1996).

Structure determination and refinement

Initial phases for the structure of L-BEK were obtained by molecular replacement, using the program AMoRe (Navaza, 1994) and the crystal structure of γ -chymotrypsin (PDB entry code 1GCD) (Harel *et al.*, 1991) as the search model. A strong unique solution was found, with correlation factors of 0.38 and 0.17 for the highest and second highest peak, respectively. Rigid body refinement followed by positional refinement using X-PLOR (Brünger, 1992) resulted in values for R and R_{free} of 43.0% and 49.2%, respectively.

The rebuilding process, using the program O (Jones & Thirup, 1986; Jones *et al.*, 1991), started by aligning the primary sequences of L-BEK and γ -chymotrypsin. The model was modified by removing the diethyl phosphate inhibitor from the chymotrypsin structure, trimming loop regions of poor sequence conservation, and then by substituting the γ -chymotrypsin residues either by alanine or by their proper counterparts in L-BEK, depending on the degree of sequence conservation. Further decreases in R and R_{free} were achieved by using the structure of thrombin (PDB entry code 1PPB) (Böde *et al.*, 1992) as a guide in regions where sequence conservation with L-BEK suggested structural similarity, building the C^{α} trace into $2F_o - F_c$ maps. At this point the

value for R_{free} dropped to 38.5%, and R decreased to 33.5%.

With the C^{α} trace in place, the model was subjected to two rounds of rebuilding guided by simulated annealing omit maps (Hodel *et al.*, 1992) in order to eliminate model bias of the initial search model with intermittent positional refinement, using the maximum likelihood target in the program CNSsolve 0.5 (Brünger *et al.*, 1998), resulting in a value for R_{free} of 33.5% that decreased to 31.5% after individual B -factor refinement. A total of 45 water molecules were added to the model and verified by inspection of the $2F_o - F_c$ electron density map. Two large spherical patches of electron density, clamped between acidic side-chains of symmetry-related molecules, were interpreted as Zn^{2+} , consistent with the presence of 20 mM zinc sulfate in the cryoprotectant solution. Their incorporation into the model led to a small but significant decrease of both R and R_{free} factors. The inhibitor Lys residue could be seen in $2F_o - F_c$ maps at an early stage of the building process, yet the remaining five residues were elusive until later in the refinement process. Eventually, residues Lys-P1 through Asp-P4 could be built in an unequivocal manner into simulated annealing omit maps, with density missing for the two N-terminal amino acid residues of the inhibitor, Asp-P5 and Val-P6. The final model comprises residues 1 through 7 of the heavy chain, residues 16 through 243 of the serine protease domain of enteropeptidase, residues P1 through P4 of the VD₄K-cm inhibitor, two Zn^{2+} and 108 water molecules. The side-chains of Lys3, Arg97 and Asn205 lacked electron density and were built as Ala. After bulk solvent correction and individual B -factor refinement, the model converged to $R = 23.4\%$ and $R_{\text{free}} = 26.9\%$ for the resolution range $30\text{--}2.3 \text{ \AA}$, using a cut-off of $F/\sigma(F) > 2.0$, with excellent stereochemistry and B -factors appropriately restrained (Table 1). There are no residues in disallowed regions of the Ramachandran plot, and only two residues in generously allowed regions.

Preparation of acetylated enteropeptidase light chain

Purified L-BEK from baculovirus (5.5 μM , 4 ml) in 0.1 M sodium phosphate (pH 7.0), was stirred on ice with 6 μl acetic anhydride added in three portions. The reaction was maintained at pH 7.0 by the dropwise addition of sodium hydroxide. After one hour, the reaction was dialyzed against 20 mM Tris-HCl (pH 7.6), 20 mM NaCl.

Enzyme kinetics

The concentration of each enteropeptidase was determined by active-site titration with p -nitrophenyl p' -guanidinobenzoate (Chase & Shaw, 1970). Kinetic parameters for cleavage of Z-Lys-SBzl were obtained as described (Green & Shaw, 1979). Assays were performed at room temperature in 1 ml of 0.1 M Tris-HCl (pH 8.0), 260 μM DTNB, and 10 μM to 500 μM Z-Lys-SBzl. Reaction was initiated by adding enzyme (0.2 to 1.6 nM) and the rate of 3-carboxy-4-nitrophenoxide production was calculated from the absorbance at 412 nm, using an extinction coefficient of $13,600 \text{ M}^{-1} \text{ cm}^{-1}$.

Kinetic parameters for the cleavage of the synthetic peptide substrate GD₄K-na were determined as described (Grant & Hermon-Taylor, 1979; Lu *et al.*, 1997). Values for K_m and k_{cat} were obtained by directly fitting to the Michaelis-Menten equation by non-linear least

squares regression. Under all assay conditions, the consumption of substrate (Z-Lys-SBzl or GD₄K-na) was <15% of the total.

Trypsinogen activation was assayed at pH 5.6 as described (Anderson *et al.*, 1977; Lu *et al.*, 1997). Assays (0.1 ml) contained 25 μ M trypsinogen, 50 mM sodium citrate (pH 5.6) at room temperature. Reaction was initiated by addition of 2 nM enteropeptidase. After ten minutes, reaction was terminated by addition of 2 μ l of 2 M HCl. To quantify the trypsin product, an equal volume of 250 μ M S-2765 in 20 mM Tris-HCl (pH 8.4), 150 mM NaCl was added and absorbance at 405 nm recorded after five minutes.

Changes in the free energy of transition state stabilization ($\Delta\Delta G_T$) were calculated from the relationship $\Delta\Delta G_T = -RT \ln (k_{cat}/K_m)_{mutant}/(k_{cat}/K_m)_{wild-type}$, where R is the gas constant, T is the absolute temperature, k_{cat} is the turnover number, and K_m is the Michaelis constant (Wilkinson *et al.*, 1983).

Inhibition by VD₄K-chloromethane

Reactions were performed in 200 μ l of 100 mM Tris-HCl (pH 8.0), VD₄K-cm (2 nM to 2 μ M) and 2 nM enteropeptidase at 22°C. At selected time intervals, 30 μ l samples were removed and added to 200 μ l of 100 mM Tris-HCl (pH 8.0), 300 μ M Z-Lys-SBzl, and 180 μ M DTNB to assay the remaining active enteropeptidase. For each concentration of inhibitor, the pseudo first-order rate constant for inactivation, k' , was determined from the relationship $\ln E = -k't + \ln E_0$, where E is the concentration of active enzyme remaining at time (t), and E_0 is the initial or total concentration of enzyme. The second-order rate constant for inactivation, k_2 , and the dissociation constant for reversible inhibitor binding, K_i , were determined from the relationship $k' = k_2[I]/([I] + K_i)$, where $[I]$ is the inhibitor concentration (Kitz & Wilson, 1962). Changes in the free energy of transition state stabilization ($\Delta\Delta G_T$) were calculated from the relationship $\Delta\Delta G_T = -RT \ln (k_2/K_i)_{mutant}/(k_2/K_i)_{wild-type}$ (Wilkinson *et al.*, 1983).

Protein Data Bank accession number

The coordinates have been deposited with the Protein Data Bank for immediate release under accession code 1ekb.

Acknowledgements

We thank Milan Kapadia for assistance in the purification of enteropeptidase variants, Dr Mark Crankshaw for performing the mass spectrometry analyses, and Dr Enrico Di Cera for advice on the refolding of recombinant proteases expressed in *E. coli*. This work was supported in part by National Institutes of Health grants DK50053 (to J.E.S.), GM54033 (to G.W.), and T32HL07088 (to D.L.). J.E.S. is an Investigator and D.L. was an Associate of the Howard Hughes Medical Institute. D.L. and K.F. contributed equally to this work.

References

Anderson, L. E., Walsh, K. A. & Neurath, H. (1977). Bovine enterokinase. Purification, specificity, and

some molecular properties. *Biochemistry*, **16**, 3354-3360.

- Baratti, J. & Maroux, S. (1976). On the catalytic and binding sites of porcine enteropeptidase. *Biochim. Biophys. Acta*, **452**, 488-496.
- Bode, W., Turk, D. & Karshikov, A. (1992). The refined 1.9-Å X-ray crystal structure of D-Phe-Pro-Arg-chloromethylketone-inhibited human alpha-thrombin: structure analysis, overall structure, electrostatic properties, detailed active-site geometry, and structure-function relationships. *Protein Sci.* **1**, 426-471.
- Bricteux-Gregoire, S., Schyns, R. & Florkin, M. (1972). Phylogeny of trypsinogen activation peptides. *Comp. Biochem. Physiol.* **42B**, 23-39.
- Brünger, A. T. (1992). *X-PLOR Version 3.1: A System for Crystallography and NMR*, Yale University Press, New Haven, CT.
- Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallog. sect. D*, **54**, 905-921.
- Carson, M. (1997). Ribbons. *Methods Enzymol.* **277**, 493-505.
- Chase, T. & Shaw, E. (1970). Titration of trypsin, plasmin, and thrombin with *p*-nitrophenyl *p*'-guanidinobenzoate HCl. *Methods Enzymol.* **19**, 20-27.
- Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallog. sect. A*, **47**, 392-400.
- Grant, D. A. W. & Hermon-Taylor, J. (1979). Hydrolysis of artificial substrates by enterokinase and trypsin and the development of a sensitive specific assay for enterokinase in serum. *Biochim. Biophys. Acta*, **567**, 207-215.
- Green, G. D. G. & Shaw, E. (1979). Thiobenzyl benzyloxycarbonyl-L-lysinate, substrate for a sensitive colorimetric assay for trypsin-like enzymes. *Anal. Biochem.* **93**, 223-236.
- Hadorn, B., Tarlow, M. J., Lloyd, J. K. & Wolff, O. H. (1969). Intestinal enterokinase deficiency. *Lancet*, **i**, 812-813.
- Harel, M., Su, C. T., Frolov, F., Ashani, Y., Silman, I. & Sussman, J. L. (1991). Refined crystal structures of "aged" and "non-aged" organophosphoryl conjugates of gamma-chymotrypsin. *J. Mol. Biol.* **221**, 909-918.
- Haworth, J. C., Gourley, B., Hadorn, B. & Sumida, C. (1971). Malabsorption and growth failure due to intestinal enterokinase deficiency. *J. Pediatr.* **78**, 481-490.
- Hodel, A., Kim, S.-H. & Brünger, A. (1992). Model bias in crystal structures. *Acta Crystallog. sect. A*, **48**, 851-858.
- Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819-822.
- Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallog. sect. A*, **47**, 110-119.
- Kalafatis, M. & Mann, K. G. (1993). Role of the membrane in the inactivation of factor Va by activated protein C. *J. Biol. Chem.* **268**, 27246-27257.

- Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W. & Sadler, J. E. (1994). Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains. *Proc. Natl Acad. Sci. USA*, **91**, 7588-7592.
- Kitz, R. & Wilson, I. B. (1962). Esters of methanesulfonic acid as irreversible inhibitors of acetylcholinesterase. *J. Biol. Chem.* **237**, 3245-3249.
- Kohnert, U., Rudolph, R., Verheijen, J. H., Weening-Verhoeff, E. J. D., Stern, A., Opitz, U., Martin, U., Lill, H., Prinz, H., Lechner, M., Kresse, G.-B., Buckel, P. & Fischer, S. (1992). Biochemical properties of the kringle 2 and protease domains are maintained in the refolded t-PA deletion variant BM 06.022. *Protein Eng.* **5**, 93-100.
- Krem, M. M. & Di Cera, E. (1998). Conserved water molecules in the specificity pocket of serine proteases and the molecular mechanism of Na⁺ binding. *Proteins: Struct. Funct. Genet.* **30**, 34-42.
- Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680-685.
- LaVallie, E. R., Rehemtulla, A., Racie, L. A., DiBlasio, E. A., Ferenz, C., Grant, K. L., Light, A. & McCoy, J. M. (1993). Cloning and functional expression of a cDNA encoding the catalytic subunit of bovine enterokinase. *J. Biol. Chem.* **268**, 23311-23317.
- Light, A. & Fonseca, P. (1984). The preparation and properties of the catalytic subunit of bovine enterokinase. *J. Biol. Chem.* **259**, 13195-13198.
- Light, A., Savithri, H. S. & Liepnieks, J. J. (1980). Specificity of bovine enterokinase toward protein substrates. *Anal. Biochem.* **106**, 199-206.
- Lu, D. & Sadler, J. E. (1998). Enteropeptidase. In *Handbook of Proteolytic Enzymes* (Barrett, A. J., Rawlings, N. D. & Woessner, J. F., Jr, eds), pp. 50-54. Academic Press Ltd, London.
- Lu, D., Yuan, X., Zheng, X. & Sadler, J. E. (1997). Bovine proenteropeptidase is activated by trypsin, and the specificity of enteropeptidase depends on the heavy chain. *J. Biol. Chem.* **272**, 31293-31300.
- Maroux, S., Baratti, J. & Desnuelle, P. (1971). Purification and specificity of porcine enterokinase. *J. Biol. Chem.* **246**, 5031-5039.
- Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., Umeyama, H., Inoue, H., Takahashi, T. & Takahashi, K. (1994). Structural characterization of porcine enteropeptidase. *J. Biol. Chem.* **269**, 19976-19982.
- Mikhailova, A. G. & Rumsh, L. D. (1999). Autolysis of bovine enteropeptidase heavy chain: evidence of fragment 118-465 involvement in trypsinogen activation. *FEBS Letters*, **442**, 226-230.
- Morrissey, J. H. (1981). Silver stain for proteins in polyacrylamide gels: a modified procedure with enhanced uniform sensitivity. *Anal. Biochem.* **117**, 307-310.
- Navaza, J. (1994). AMoRe: an automated package for molecular replacement. *Acta Crystallog. sect. A*, **50**, 157-163.
- Nelson, R. M. & Long, G. L. (1989). A general method of site-specific mutagenesis using a modification of the *Thermus aquaticus* polymerase chain reaction. *Anal. Biochem.* **180**, 147-151.
- Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281-296.
- Otwinowski, Z. & Minor, W. (1996). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307-326.
- Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411-2423.
- Pavlov, I. P. (1902). *The Work of the Digestive Glands*, Trans. Charles Griffin & Co. W. H. Thompson, London.
- Schechter, I. & Berger, A. (1967). On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **27**, 157-162.
- Sheehan, J. P. & Sadler, J. E. (1994). Molecular mapping of the heparin-binding exosite of thrombin. *Proc. Natl Acad. Sci. USA*, **91**, 5518-5522.
- Vijayalakshmi, J., Padmanabhan, K. P., Mann, K. G. & Tulinsky, A. (1994). The isomorphous structures of prethrombin2, hirugen-, and PPACK-thrombin: changes accompanying activation and exosite binding to thrombin. *Protein Sci.* **3**, 2254-2271.
- Wilkinson, A. J., Fersht, A. R., Blow, D. M. & Winter, G. (1983). Site-directed mutagenesis as a probe of enzyme structure and catalysis: tyrosyl-tRNA synthetase cysteine-35 to glycine-35 mutation. *Biochemistry*, **22**, 3581-3586.
- Wu, Q., Sheehan, J. P., Tsiang, M., Lentz, S. R., Birktoft, J. J. & Sadler, J. E. (1991). Single amino acid substitutions dissociate fibrinogen-clotting and thrombomodulin-binding activities of human thrombin. *Proc. Natl Acad. Sci. USA*, **88**, 6775-6779.
- Yahagi, N., Ichinose, M., Matsushima, M., Matsubara, Y., Miki, K., Kurokawa, K., Fukamachi, H., Tashiro, K., Shiokawa, K., Kageyama, T., Takahashi, T., Inoue, H. & Takahashi, K. (1996). Complementary DNA cloning and sequencing of rat enteropeptidase and tissue distribution of its mRNA. *Biochem. Biophys. Res. Commun.* **219**, 806-812.
- Yuan, X., Zheng, X. L., Lu, D. S., Rubin, D. C., Pung, C. Y. M. & Sadler, J. E. (1998). Structure of murine enterokinase (enteropeptidase) and expression in small intestine during development. *Am. J. Physiol.* **37**, G342-G349.

Edited by R. Huber

(Received 27 May 1999; received in revised form 23 July 1999; accepted 26 July 1999)



Exhibit 22

Structural Characterization of Porcine Enteropeptidase*

(Received for publication, March 8, 1994, and in revised form, April 11, 1994)

Masashi Matsushima†§, Masao Ichinose†, Naohisa Yahagi†, Nobuyuki Kakei†, Shinko Tsukada†, Kazumasa Miki†, Kiyoshi Kurokawa†, Kosuke Tashiro†, Koichiro Shiokawa†, Kazuko Shinomiya†, Hideaki Umeyama†, Hideshi Inoue§, Takayuki Takahashi§, and Kenji Takahashi§**

From the †First Department of Internal Medicine, Faculty of Medicine and the Departments of §Biophysics and Biochemistry and ‡Zoology, Faculty of Science, University of Tokyo, Tokyo 113 and the §School of Pharmaceutical Sciences, Kitasato University, Tokyo 108, Japan

Enteropeptidase (EC 3.4.21.9) is a key enzyme in the intestinal digestion cascade responsible for the conversion of trypsinogen to trypsin, which then activates various pancreatic zymogens. In order to structurally characterize the enzyme, we purified the enzyme from porcine duodenal mucosa and showed that it consists of three polypeptide chains, which we named "mini" chain (M chain), light chain (L chain), and heavy chain (H chain) in order of increasing molecular size. Based on their NH₂-terminal sequences, a cDNA clone for porcine enteropeptidase was isolated and analyzed. The clone was 3597 base pairs long, which encoded 1034 amino acid residues of a single-chain precursor form of enteropeptidase. The precursor contained an additional NH₂-terminal 51-residue sequence including a putative internal signal sequence, followed by the M chain (66 residues), the H chain (682 residues), and the L chain (235 residues) in that order. The H chain had regions partially homologous in sequence with low density lipoprotein receptor and complement components. On the other hand, the L chain was highly homologous with the catalytic domains of trypsin-like serine proteinases. The structural model of the L chain suggests that the sequence, Arg⁸⁶³-Arg-Arg-Lys⁸⁶⁸, is probably involved in the unique substrate specificity of the enzyme, preferring acidic amino acid residues at the P₁-P₂ sites.

Enteropeptidase (enterokinase, EC 3.4.21.9) is well known and physiologically the only enzyme capable of converting trypsinogen to trypsin (1). Trypsin thus produced then converts various pancreatic zymogens including trypsinogen itself to their corresponding active enzymes. Therefore, enteropeptidase has been recognized to play a key role in regulating intestinal protein digestion. Indeed, patients with primary enteropeptidase deficiency, a genetic disorder with no or little enteropeptidase activity in the duodenum, have been reported to suffer from malabsorption and malnutrition, particularly in infancy, and need to take drugs containing a pancreatic enzyme mixture for recovery (2).

Because of its physiological importance, there have been a number of studies on the purification and characterization of enteropeptidase from various species (3-9). These studies have

shown that the enzyme is classified as a trypsin-like serine proteinase having strict specificity toward substrates with a basic amino acid residue at the P₁ site¹ and acidic residues at the P₂-P₃ sites as expected from the NH₂-terminal amino acid sequence (Val¹-Asp-Asp-Asp-Asp-Lys⁶) of bovine trypsinogen. In contrast, structural information on the enzyme is still limited. Its molecular weight thus far reported ranges from 150,000 to 300,000, depending on the difference in species. In addition, the number of constituent polypeptide chains has been reported differently; the enzyme was reported to be composed of two chains in pig (4) and cow (7, 9) and three chains in human (10). Available data indicate that in all cases the smaller polypeptide chain, called the light chain, is a catalytic chain (4, 10, 11), but the precise chain composition is not yet as clear. This is largely due to lack of information on the complete amino acid sequence of enteropeptidase, although the bovine light chain sequence has been reported very recently by LaVallie *et al.* (12).

We have recently established a purification procedure for enteropeptidase from porcine duodenal mucosa and found that, unlike the previous data (4), the enzyme consists of three different polypeptide chains, i.e. "mini" (M),² light (L), and heavy (H) chains. Furthermore, we have cloned and analyzed a cDNA coding for the protein and deduced its complete amino acid sequence. The results clearly indicate that enteropeptidase is synthesized as a single-chain precursor protein and then is processed to the mature enzyme. In this paper, we describe these results and discuss the substrate specificity of the enzyme based on the three-dimensional structure constructed by computer modeling.

MATERIALS AND METHODS

Determination of Protein Concentration—Protein concentration was estimated colorimetrically by using a protein assay kit (Bio-Rad) and mouse IgG as the standard (13).

Enzyme Purification—Enzyme activity was assayed essentially according to Liepnies and Light (7) with some modification. The purification procedure will be described in detail elsewhere. In brief, the mucosa was obtained from 40 porcine duodena by squeezing them with the fingers in 20 mM Tris-HCl (pH 8.0), and the crude extract was obtained from the mucosa by solubilizing with 1% sodium deoxycholate followed by centrifugation. The enzyme was purified from the extract by four steps of chromatography on columns of DE52 (5.4 × 40 cm, Whatman), Butyl Toyopearl 650S (2 × 20 cm, prepacked, Tosoh), Sephacryl S-300 (3.6 × 90 cm, Pharmacia Biotech Inc.), and benzamidine-Sepharose (0.9 × 25 cm, Pharmacia). The enzyme fractions obtained from the last column were pooled, concentrated, and used for further experiments.

* This work was supported in part by grants-in-aid for scientific research from the Ministry of Education, Science and Culture of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) D30799.

** To whom correspondence and reprint requests should be addressed. Tel.: 81-3-5689-5607; Fax: 81-3-5802-2041.

¹ The nomenclature is according to Berger and Schechter (60).

² The abbreviations used are: M chain, "mini" chain; L chain, light chain; H chain, heavy chain; LDL, low density lipoprotein; PAGE, polyacrylamide gel electrophoresis.

TABLE I
Purification of porcine duodenal enteropeptidase
EKU is defined as nanomoles of trypsin produced in 30 min at 37 °C.

Step	Total protein mg	Total activity EKU	Specific activity EKU/mg protein	Yield %	Purification -fold
Crude extract	4,730	157,000	33.2	100	1
DE52	304	58,300	192	37.1	5.8
Butyl Tboyparyl	35.2	27,500	720	17.5	21.7
Sephacryl S-300	2.94	13,300	4,530	8.5	136
Benzamidine-Sepharose	0.42	10,000	24,200	6.4	729

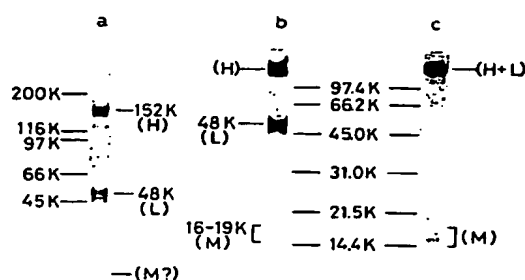


FIG. 1. SDS-PAGE patterns of the purified enzyme. a, under reducing conditions using a gradient gel of 4–20%; b, under reducing conditions using a gradient gel of 15–25%; c, under nonreducing conditions using a gradient gel of 15–25%. Approximately 30 µg of the enzyme was applied to each lane.

Polyacrylamide Gel Electrophoresis—Polyacrylamide gel electrophoresis (PAGE) was performed essentially according to Laemmli (14) using SDS-PAGE plate 4/20 and Multigel 15/25 (Daiichi, Tokyo).

NH₂-terminal Amino Acid Sequence Analysis—The purified enzyme sample was subjected to SDS-PAGE using 4–20 or 15–25% gradient gels, and the separated polypeptides were transferred to Immobilon P (Millipore) or Immobilon P⁵⁴ (Millipore) essentially according to LeGendre and Matsudaira (15). The proteins on the membranes were analyzed with an automated protein sequencer (model 477A, Applied Biosystems) on-line to a phenylthiohydantoin-derivative analyzer (model 120A, Applied Biosystems).

cDNA Cloning and Analyses—The total RNA was extracted from freshly resected porcine duodenal mucosa by the guanidium isothiocyanate method and purified by CsCl density gradient ultracentrifugation (16). The poly(A) RNA was isolated using Oligotex dT-30 super (Takara). Complementary double-stranded DNA was synthesized using a cDNA synthesis system plus (Amersham Corp.) from 5 µg of the poly(A) RNA as a template with oligo(dT) or random hexanucleotide as a primer (17). The cDNA libraries were constructed using a cDNA cloning system (Amersham Corp.), except that λZAP II/EcoRI vector (Stratagene) was used. A 53-mer oligonucleotide described under "Results" was synthesized by Sawadaya Technology (Tokyo). The probe was labeled at the 5'-end using [γ -³²P]ATP (6000 Ci/mmol, Amersham Corp.) and a Megalabel labeling kit (Amersham Corp.). The DNA fragment probe was labeled by the multiprimer method using (α -³²P)dCTP (3000 Ci/mmol, Amersham Corp.) and a Megaprimer labeling kit (Amersham Corp.). The transfer membrane used was Hybond N (Amersham Corp.), and the conditions of transfer, fixation, prehybridization, hybridization, and wash were essentially according to the manufacturer. For the 53-mer oligonucleotide probe, 45 °C was adopted as the temperature of prehybridization and hybridization, and 2 × SSC and 0.1% SDS at 50 °C as the stringent wash conditions. The cloned cDNA in the vector was automatically subcloned to pBluescript phagemid, and double-stranded DNA in the phagemid was used as a template for DNA sequencing. DNA sequencing was performed by the dideoxy chain termination method (18) using a Taq dye primer sequencing kit (Applied Biosystems), a thermal cycler (model PJ 480, Perkin-Elmer), and a DNA sequencer (model 370A, Applied Biosystems).

Computer Modeling of Three-dimensional Structure of L Chain—A homology search for the L chain was performed in the Brookhaven Protein Data Bank by the multiple alignment system for protein se-

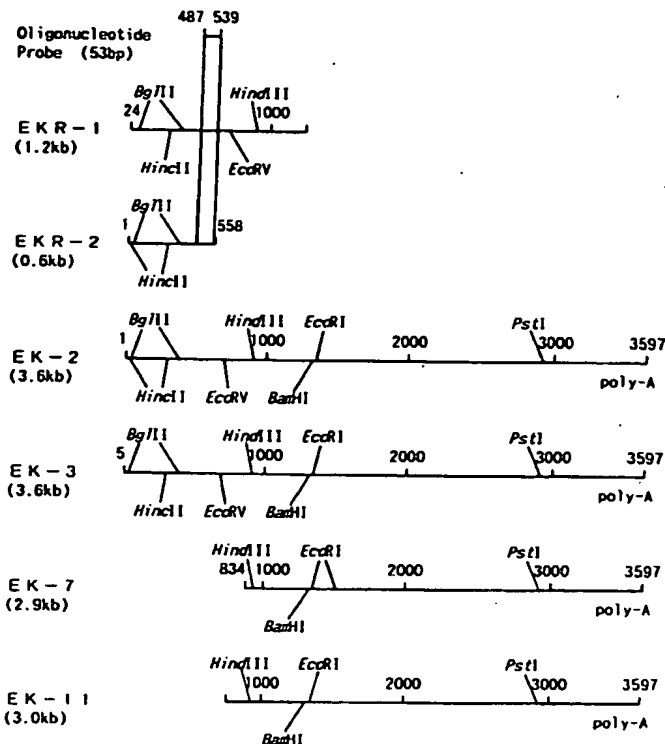


FIG. 2. Restriction enzyme mapping of the cDNA clones. The base pair numbers are according to the numbering of the longest clone, EK-2. EK-1 and -2 were positive clones in the random-primed cDNA library, while EK-2, -3, -7, and -11 were positive in the oligo(dT)-primed library. All clones had the same map except for an EcoRI site in EK-7.

quences (62). Comparing the sequences of the 28 most homologous proteins of known three-dimensional structure with that of the porcine L chain, the L chain was divided into 13 parts so that each segment had a similar deletion and insertion profile. For each segment, one protein was selected from the homology list so as to minimize insertion and deletion and to maximize identity. Thus, a chimeric reference protein was constructed that was composed of the following segments: 1HNE (human neutrophil elastase) for positions 800–814, 815–825, and 839–856; 1DWB (human thrombin) for 826–838 and 869–892; 3RP2 (A chain, rat mast cell protease II) for 857–868; 4CHA (A chain, bovine α -chymotrypsin) for 893–930, 988–1003, and 1018–1034; 3EST (porcine pancreatic elastase) for 931–944; 1SGT (*Streptomyces griseus* trypsin) for 945–971; and 1TLD (bovine β -trypsin) for 972–987 and 1004–1017. Gly⁸⁴⁴ and Arg⁸⁴⁵ were inserted into the reference protein 1HNE by using the coordinates of the main chain of Gln-Arg of Leu²²²-Tyr-Gln-Gln-Arg-Asp-Val-Asn²²⁹ of 6TIM (triose-phosphate isomerase). The three-dimensional modeling of the L chain was performed using the chimeric protein as a reference protein according to Kajihara *et al.* (19). Modeling of the complex of the L chain and Val-(Asp)₄-Lys was also performed with the above structural model as a base protein using the coordinates of the main chain of Lys¹³-Pro-Ala-Cys-Thr-Leu¹⁸ of the inhibitor part in 3SGB in protein data bank code (proteinase B from *S. griseus* complexed with the third chain of turkey ovomucoid inhibitor) for the initial arrangement of the hexapeptide, essentially according to the same method.

RESULTS

Purification and Structural Characterization of Porcine Enteropeptidase—From 40 porcine duodena, 0.42 mg of the purified enzyme was obtained in a 6.4% yield with 729-fold purification (Table I). The molecular weight of the enzyme was estimated to be approximately 200,000 by gel filtration (data not shown). As shown in Fig. 1a, SDS-PAGE using a gradient gel (4–20%) under reducing conditions gave two polypeptide

[illegible]

The NH₂-terminal amino acid sequences of the H and L chains of the enzyme were shown to be SVIVFDLLFAQWVS-DENIKEELIQGIEA (29 residues) and IVGGXDSREGAXPKV-VALYYNGQLXLXGASLV (31 residues), respectively. For the H chain, the analyses of the three bands electrophoretically separated on SDS-PAGE resulted in the same sequence of LGKS-HEARGTMKTTXGVTYNPNL (23 residues). The molar ratio of the H, L, and M chains in the enzyme estimated from the amounts of phenylthiohydantoin-derivatives obtained by NH₂-

	Molecular weight, $\times 10^3$		Number of potential asparagine-linked glycosylation sites
	Calculated	Measured by SDS-PAGE	
M chain	7.5	16-19	1
H chain	75.4	152	17
L chain	26.4	48	4

Isolation and Characterization of Porcine Enteropeptidase cDNA Clones—Based on part of the NH₂-terminal sequence of the H chain (Phe¹⁰ to Ile³⁷), we designed a 53-mer oligonucleotide probe including 16 inosines, 8-fold redundant and comple-

FIG. 4. Comparison of the amino acid sequence of the catalytic chain of enteropeptidase with those of other serine proteinases. The catalytic chain sequence of porcine enteropeptidase is compared with those of bovine enteropeptidase (12), human hepsin (21), human plasma kallikrein (22), human factor XIa (45), dog trypsin (46), bovine trypsin (47), bovine chymotrypsin (48-51), and porcine elastase (52). Residues are expressed in one-letter code. "*" indicates the same residue with porcine enteropeptidase; "-" indicates deletion inserted to optimize the homology. Residues in *white letters* are the conserved catalytic triad, His, Asp, and Ser. The percentages of identity with porcine enteropeptidase are listed at the ends of the sequences.

Nucleotide and Deduced Amino Acid Sequences of cDNA Clone EK-2—The nucleotide and the deduced amino acid sequences of EK-2 are shown in Fig. 3. The cDNA clone was 3597 base pairs long. It had a polyadenylation signal at the 3559 base pair position and poly(A) at the 3'-end. The first ATG met the criteria for an initiator codon in eukaryotes (20). Assuming this codon to be the initiator, the open reading frame was 3102 base pairs long, and thus the deduced amino acid sequence was composed of 1034 residues. The boxed sequence from positions 19 to 43 was the most hydrophobic domain in the sequence. The NH₂-terminal sequences of the M, H, and L chains were deduced to start at positions 52, 118, and 800, respectively. Thus, the enzyme is thought to be originally synthesized as a single-chain precursor ($M_r = 114,763$). Assuming that no more processing occurs in the COOH-terminal region of each chain, the

A homology search for the deduced amino acid sequence by the FASTA program in the PIR protein data base revealed that the catalytic (L) chain is homologous with those of trypsin- and chymotrypsin-like serine proteinases (Fig. 4). Human hepsin (21) and plasma kallikrein (22) showed over 40% identity. The bovine enzyme (12) was 89.8% identical with the porcine enzyme. On the other hand, the H chain had interesting homologies in limited regions of certain proteins. The sequences at positions 195–236 and 654–692, homologous with each other, were homologous with those in complement C9 (23), low density lipoprotein (LDL) receptor (24), etc. (Fig. 5a). The sequences at positions 240–353 and 539–653 are also homologous with each other and were homologous with those in dorsal-ventral patterning protein (25), complements C1r (26) and C1s (27), etc. (Fig. 5b). The sequence at positions 772–788 was homologous with those in factor X (28), protein C (29), hepsin (21), etc. (Fig. 5c).

Three-dimensional Structure of L Chain of Porcine Enteropeptidase as Deduced by Computer Modeling—Three-dimensional structural modeling of the complex of the catalytic chain and the NH₂ terminus of bovine trypsinogen, Val¹-Asp-Asp-Asp-Lys⁶, was performed using the chimeric reference protein, which was 38.7% identical with the L chain with a 2-residue insertion in the fourth segment: The resulting model³ is shown in Fig. 6a. The mode of binding of the NH₂-terminal

³ The coordinate data of the model may be presented on request.

a C9/LDL-receptor type region

Enteropeptidase (195-238)
(654-692)

Consensus sequences
LDL-receptor
Terminal complement components
LDL-receptor related protein
Perlecan
GP-330

VSIECLPGSRFCADALNCIAVDLFCDGELDCPDGSDSDSKIC
IPERCKEDMFCEN-GEVLLVDLCDFGFSKCKDESEAH--C
...TC...EF...G...CI...W...CD...DC...DGSDE...C
E...CG...DFQC...T...GRCKRRRL...CNGD...DCED...SDDD...C
...C...F...C...RCIP...W...CDG...DC...D...SDE...C
P-PC-P-EF-C...C...CD...DC...D...SDE...C
C...F...C...CI...C...CDG...DC...DGSDE...C

b C1r/s type region

Enteropeptidase (240-353)
(539-653)

Consensus sequence
(C1r/s, DVPP, BMP-1)

CDGKFLITESSSF-DAAYPKL-SEASVVCWILRYNCGLSIELNFSY--RNTYSM-----
CGGPFELWEPHTTF-TSMNPPNH-YFNQAFQVNLMAQKGNICLDFEE--FDLENIA-----
CG...L...T...G...I...S...P...Y...Y...C...W...T...A...G...V...L...F...FDLE...
V
--DVLNITYEGVGSKILRASLWM--NPGTIRIFSNQVTVTLIESDENQYL--GFNATYTAENSTE
--DVEIRGDEEDSLLLA-VYTG--PGPVEDVFSTTNRMTMLFIITDALTKG--GFKANFTTGYYHLG
C-YD-L-T-G...G...C...R...P...D...T...N...L...L...F...SD...S...GF...A...
V S S M Y T T L

c Carboxyl-terminal region of the non-catalytic chain

Enteropeptidase (772-788)
Hepsin (140-154)
Factor X (111-133)
Protein C (120-142)

Q---FEDSLILLCNHKS---CG
QPRGRFLAAI---CGD---CG
CARGYTLADNGKACIPTGPYPCG
GAPGYKLGDDLLCHPAVKFPCG

FIG. 5. Comparison of partial sequences of the H chain with those of homologous regions in other proteins. *a*, the cysteine-rich sequence repeats are compared with the consensus sequences of human LDL receptor (24); human terminal complement components C7 (53), C8 α (54), C8 β (55), and C9 (23); human LDL receptor-related protein (56); human perlecan (57); and rat GP-330 (58). The residues identical in at least six sequences are boxed. *b*, C1r/s type sequences are compared with the consensus sequence (25) among the sequences of human complement components C1r (26) and C1s (27), *Drosophila* dorsal-ventral patterning protein (DVPP) (25), and bone morphogenetic protein-1 (BMP-1) (59). The residues identical between the enteropeptidase sequences and the consensus sequence are boxed. *c*, the sequence near the carboxyl-terminal end of the H chain is compared with those of the corresponding regions of human hepsin (21), human factor X (42), and human protein C (29). The residues identical in the four sequences are boxed. In *a*, *b*, and *c*, the values in parentheses indicate residue numbers; "--", a deletion inserted to optimize the homology; ".", a non-consensus residue.

hexapeptide of bovine trypsinogen with the active site region of the catalytic chain is also shown (Fig. 6b).

DISCUSSION

The mature three-chain enzyme is thought to be generated by peptide bond cleavages from the single-chain precursor in which the three chains are aligned in the order M, H, and L chains, starting from the NH₂ terminus. Previously, the porcine enzyme was reported to be composed of two chains, an H chain ($M_r = 134,000$) and an L chain ($M_r = 62,000$) (4). On the other hand, the human enzyme was reported to be a three-chain enzyme (10). Two of the human chains have molecular weights of 140,000 and 54,000, comparable with those of the H and L chains of the porcine enzyme, respectively, but the third polypeptide ($M_r = 120,000$) of the human enzyme is much larger than the porcine M chain ($M_r = 16,000-19,000$). Thus, the M chain appears to be a newly identified component of the enzyme, although it is not clear at present whether the M chain is essential for the function of enteropeptidase.

The predicted amino acid sequence of the porcine enteropeptidase precursor contained a 51-residue peptide sequence, which is missing in the purified mature enzyme. This peptide contains a very hydrophobic segment (from Val¹⁹ to Ile⁴⁹) long enough to span the membranes. Since the precursor protein does not appear to have any other membrane-spanning segment or typical signal sequence, this hydrophobic segment presumably serves as an internal signal sequence (30-32) and keeps the enzyme bound to membranes. Enteropeptidase is localized to the brush border membranes of the duodenum and upper intestine (33, 34) in such a manner that its catalytic

domain can freely contact extracellular trypsinogen. Therefore, the NH₂-terminal region should reside on the cytoplasmic side and the COOH-terminal region on the outside of the cell. Thus, enteropeptidase is apparently a Type II⁴ integral membrane protein. The NH₂-terminal positively charged residue(s) flanking the internal signal sequence is known to be an important part of a dominantly acting retention signal to create the Type II orientation (35). The NH₂-terminal 51-residue peptide apparently meets the above structural requirements.

As schematically shown in Fig. 7, the purified porcine enzyme obviously resulted from proteolytic cleavages at three sites. Cleavage at Ala⁵¹-Leu⁵² produces the enzyme dissociated from the membranes. Interestingly, Toyoda *et al.* (36) reported that elastase could release enteropeptidase activity from the brush border membranes. The peptide bond cleavage at Ala⁵¹-Leu⁵² is compatible with the substrate specificity of elastase. Therefore, elastase may be responsible for the cleavage. In addition, other proteinases cleaving Gly¹⁰⁷-Ser¹⁰⁸ and Lys⁷⁹⁹-Ile⁸⁰⁰ must be present, although no information about them is available at present.

The H chain has a Ser/Thr-rich sequence at positions 172-187, comprising 12 residues of Ser/Thr. Such Ser/Thr-rich regions, which have been found in glycoprotein A (37), LDL receptor (38), sucrase-isomaltase (39), aminopeptidase N (40), etc., are documented to be potential O-linked glycosylation sites. Indeed, polyclonal antibodies against human enteropeptidase were reported to cross-react with type A blood antigen (10), indicating the presence of O-linked oligosaccharide(s) in the

⁴ The nomenclature is according to von Heijne and Gavel (61).

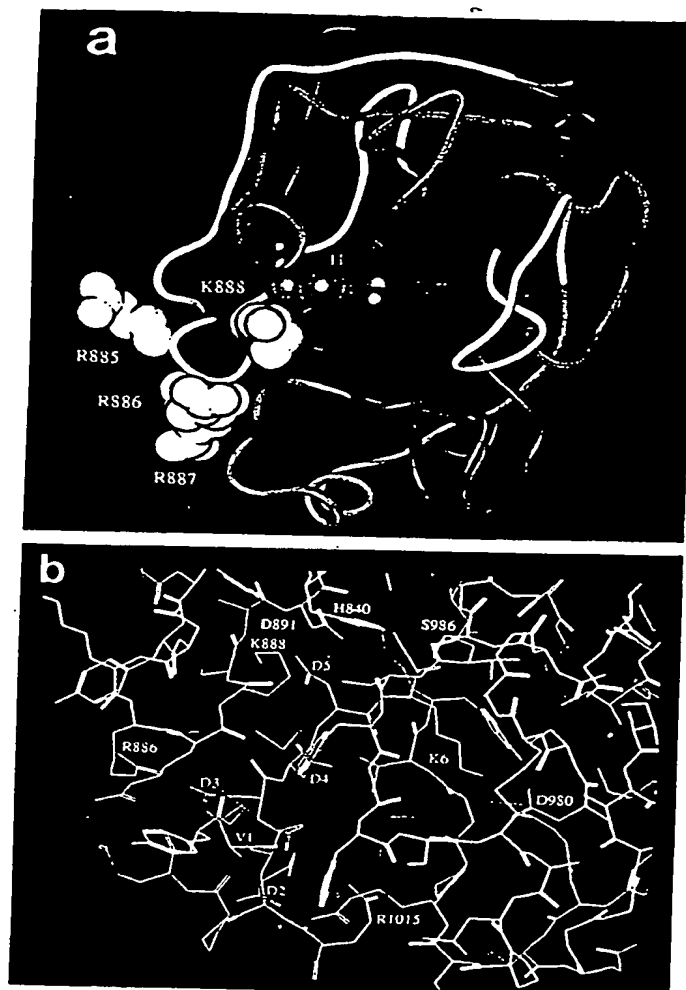


Fig. 6. The three-dimensional structure of the L chain of porcine enteropeptidase constructed by computer modeling. *a*, the tube model of the main chain. Segments in the reference chimera protein derived from 3RP2, 1TLD, 1DWB, 4CHA, 1SGT, 1HNE, and 3EST are colored in red, green, yellow, blue, magenta, cyan, and white, respectively. The side chains in the basic amino acid cluster, Arg⁸⁸⁵-Arg-Arg-Lys⁸⁸⁶, are shown with the Corey-Pauling-Koltun models colored in yellow, and those of the active site His⁸⁴⁰ and Ser⁸⁸⁶ in cyan and green, respectively. The ribbon model colored in red shows the main chain of part of the substrate, Val-Asp-Asp-Asp-Lys. *b*, the stick model of the enzyme interacting with part of the substrate, Val-Asp-Asp-Asp-Lys. The substrate part is shown with the red stick model. The side chains of the catalytic triad of Asp⁸⁹¹, His⁸⁴⁰, and Ser⁸⁸⁶ of the enzyme are shown with the yellow stick model, and those of the amino acid residues of the enzyme interacting with the substrate are shown with the blue stick model. The calculated distances for the two hydrogen bonds, His⁸⁴⁰N^H...Asp⁸⁹¹O^H and His⁸⁴⁰N^H...Ser⁸⁸⁶O^H, and the ionic pair, Arg⁸⁸⁵N^H...Glu⁸⁸²O^H, are 2.74, 3.08, and 2.65 Å, respectively. Those for the ionic pairs between the enzyme and the substrate trypsinogen (Arg¹⁰¹⁵N^H...Asp⁸⁹¹O^H, Arg⁸⁸⁶N^H...Asp⁸⁹¹O^H, and Lys⁸⁸⁷N^H...Asp⁸⁹¹O^H) and the hydrogen bonds between the enzyme and substrate main chain atoms (Tyr¹⁰⁰⁸N-Asp⁸⁹¹O, Gly¹⁰⁰⁷N-Asp⁸⁹¹O, and Gly⁸⁸⁴N-Lys⁸⁹⁰O) are 2.66, 2.76, and 2.51 Å and 2.76, 2.77, and 2.69 Å, respectively.

enzyme. Thus, the Ser/Thr-rich segment in the H chain is presumably the region of O-linked carbohydrate attachment. In addition, 22 potential N-linked glycosylation sites are seen in the enzyme, in accord with the previous findings that the enzyme is heavily glycosylated (4, 6, 7). From the present study,

the carbohydrate content of porcine enteropeptidase is estimated to be as much as 50% of the total weight.

Two sets of repeating sequences are present in the H chain. We found two tandem repeats of 38 amino acids (about 30% identity) including 6 conserved cysteine residues (Fig. 5*a*). Although the locations of the disulfide bonds in enteropeptidase have not been determined, these 6 cysteine residues are likely to form three intrachain disulfide bonds within each of the two repeats. They are homologous with certain regions in some terminal complement components such as C9 (23), LDL receptor (24), etc. The homologous seven repeating sequences in LDL receptor are thought to be the sites for interaction with apolipoproteins (38). Besides, polymeric complement C9 has recently been reported to have affinity with apolipoproteins (41). By analogy, the cysteine-containing repeats in enteropeptidase may also be the sites of interaction with other proteins such as apolipoproteins. As shown in Fig. 5*b*, the H chain contains another two segments with internal homology (about 25% identity), resembling partial sequences of complement components C1r (26) and C1s (27), etc. At present, the role of this C1r/s-type region in the enteropeptidase H chain is not known. In addition, a region near the COOH-terminal end of the H chain shows low but detectable sequence homology with the corresponding regions of the non-catalytic chains of some other serine proteinases (Fig. 5*c*). In protein C (29) and factor X (42), proteolytic cleavages in the activation process are known to occur at mono- or dibasic sites between these regions and the NH₂ termini of the catalytic chains. By analogy, the enteropeptidase precursor may be cleaved at the dibasic site Lys⁷⁸⁹-Lys⁷⁹⁰ at first and then activated by the cleavage at the NH₂ terminus of the L chain.

On the other hand, the L chain is highly homologous with the catalytic chains of other serine proteinases (Fig. 4). The three-dimensional structural model of the L chain indicates that the catalytic triad, His⁸⁴⁰, Asp⁸⁹¹, and Ser⁸⁸⁶, and the S₁' pocket are situated essentially in the same manner as in trypsin (43). Moreover, in the S₁ pocket, Asp⁸⁹⁰ positioned at its bottom and Gly¹⁰⁰⁷ and Gly¹⁰¹⁷ at its neck are also conserved in enteropeptidase, indicating that it is a typical trypsin-like serine proteinase. Since enteropeptidase has a strict specificity toward substrates with acidic amino acid residues at the P₂-P₃ sites, the presence of additional sites (S₂-S₃) for substrate side chain binding has been postulated (3, 8). Lysine residue(s) has been suggested to be important to the substrate specificity of porcine enteropeptidase by a chemical modification study (44). According to the present structural model of the porcine L chain including the NH₂-terminal hexapeptide (Val¹-Asp-Asp-Asp-Lys⁶) of bovine trypsinogen (Fig. 6*b*), the basic cluster sequence, Arg⁸⁸⁵-Arg-Arg-Lys⁸⁸⁶, unique to enteropeptidase among the family of serine proteinases (Fig. 4), appears to make a turn structure adjacent to the S₁ pocket and interact with Asp²-Asp-Asp⁵ of trypsinogen through three strong salt bridges: Arg¹⁰¹⁵ versus Asp², Arg⁸⁸⁶ versus Asp³, and Lys⁸⁸⁷ versus Asp⁵. This is consistent with the previous results indicating that an acidic amino acid at the P₂ site in the substrate is essential and that those at the P₃-P₅ sites are beneficial for the cleavage (3, 8). In the bovine L chain, the residue corresponding to Arg⁸⁸⁵ is substituted with Lys (12), but the substitution does not seem to cause any significant effect on the interaction with the substrates. Moreover, Arg⁸⁸⁷ makes an ion pair with Glu⁸⁸². The carboxyl group of Asp⁴ of the peptide does not interact with the enzyme in this model but may form an ion pair with the side chain of Lys¹⁷⁶ of bovine trypsinogen as judged from a three-dimensional structure model (data not shown). Further, the main chain atoms, Asp²O, Asp³O, and Lys⁶O of the peptide

Structure of Porcine Enteropeptidase

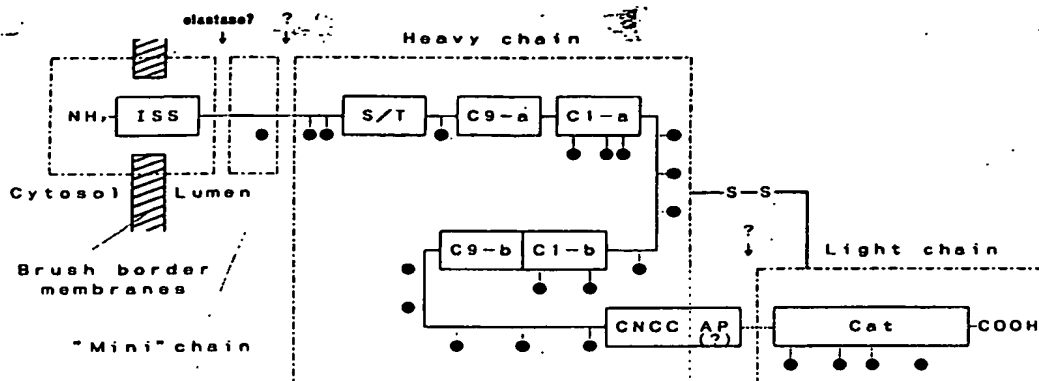


Fig. 7. The gross structure of the precursor form of porcine enteropeptidase, the sites of proteolytic processing, and potential asparagine-linked glycosylation sites. ISS, putative internal signal sequence; S/T, Ser/Thr-rich sequence; C9-a and -b, repeating sequences homologous with part of the sequences of complement C9/LDL receptor; C1-a and -b, repeating sequences homologous with part of the sequences of complement C1r/s; CNCC, sequence near the COOH-terminal region of the H chain homologous with those of the noncatalytic chains of two-chain serine proteinases such as factor X and protein C; AP, putative activation peptide; Cat, catalytic domain. Closed circles indicate potential asparagine-linked glycosylation sites. Vertical arrows indicate proteolytic processing sites.

substrate form three hydrogen bonds with the atoms, Tyr¹⁰⁰⁸N, Gly¹⁰⁰⁷N, and Gly⁹⁸⁴N of the enzyme, respectively. Thus, the unique substrate specificity of enteropeptidase can be explained clearly.

Acknowledgments—We thank Dr. S. B. P. Athauda, Dr. Y. Tamaoue, and Y. Tsuchiya for valuable discussion of this work; Y. Sakurai for NH₂-terminal amino acid sequence analysis and kind advice on measurement of the enzyme activity; and Dr. H. Komooka and Dr. K. Kamiya for the computer programs used in deducing the three-dimensional structure of the catalytic chain.

REFERENCES

- Light, A., and Janska, H. (1989) *Trends Biochem. Sci.* 14, 110–112
- Ghishan, F. K., Lee, P. C., Lebenthal, E., Johnson, P., Bradley, C. A., and Greene, H. L. (1983) *Gastroenterology* 85, 727–731
- Maroux, S., Baratti, J., and Desnuelle, P. (1971) *J. Biol. Chem.* 246, 5031–5039
- Baratti, J., Maroux, S., Louvard, D., and Desnuelle, P. (1973) *Biochim. Biophys. Acta* 315, 147–161
- Grant, D. A. W., and Hermon-Taylor, J. (1975) *Biochem. J.* 147, 363–366
- Grant, D. A. W., and Hermon-Taylor, J. (1976) *Biochem. J.* 155, 243–254
- Liepnies, J. J., and Light, A. (1979) *J. Biol. Chem.* 254, 1677–1683
- Light, A., Savithri, H. S., and Liepnies, J. J. (1980) *Anal. Biochem.* 106, 199–206
- Fonseca, P., and Light, A. (1983) *J. Biol. Chem.* 258, 14516–14520
- Magge, A. I., Grant, D. A. W., and Hermon-Taylor, J. (1981) *Clin. Chim. Acta* 115, 241–254
- Light, A., and Fonseca, P. (1984) *J. Biol. Chem.* 259, 13195–13198
- LaValle, E. R., Rehmtulla, A., Racio, L. A., DiBlasio, E. A., Ferenz, C., Grant, K. L., Light, A., and McCoy, J. M. (1993) *J. Biol. Chem.* 268, 23311–23317
- Bradford, M. M. (1976) *Anal. Biochem.* 72, 248–254
- Laemmli, U. K. (1970) *Nature* 227, 680–685
- LeGendre, N., and Matsudaira, P. (1989) in *A Practical Guide to Protein and Peptide Purification for Microsequencing* (Matsudaira, P., ed.), pp. 52–72. Academic Press Inc., San Diego, CA
- Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W. J., and Goodman, H. M. (1977) *Science* 196, 1313–1319
- Gubler, U., and Hoffman, B. J. (1983) *Gene (Amst.)* 25, 263–269
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467
- Kajihara, A., Komooka, H., Kamiya, K., and Uneyama, H. (1993) *Protein Eng.* 6, 615–620
- Kozak, M. (1984) *Nucleic Acids Res.* 12, 857–872
- Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* 27, 1067–1074
- Chung, D. W., Fujisawa, K., McMullen, B. A., and Davie, E. W. (1986) *Biochemistry* 25, 2410–2417
- DiScipio, R. G., Gehring, M. R., Podack, E. R., Kan, C. C., Hugli, T. E., and Fey, G. H. (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 7298–7302
- Südhof, T. C., Goldstein, J. L., Brown, M. S., and Russell, D. W. (1985) *Science* 228, 815–822
- Shimell, M. J., Ferguson, E. L., Childs, S. R., and O'Connor, M. B. (1991) *Cell* 67, 469–481
- Journet, A., and Tsai, M. (1986) *Biochem. J.* 240, 783–787
- Mackianon, C. M., Carter, P. E., Smyth, S. J., Dunbar, B., and Fothergill, J. E. (1987) *Eur. J. Biochem.* 169, 547–555
- McMullen, B. A., Fujisawa, K., Kisiel, W., Sasagawa, T., Howald, W. N., Kwa, E. Y., and Weinstein, B. (1983) *Biochemistry* 22, 2875–2884
- Foster, D., and Davie, E. W. (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 4766–4770
- Bos, T. J., Davis, A. R., and Nayak, D. P. (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 2327–2331
- Spies, M., and Lodish, H. F. (1986) *Cell* 44, 177–185
- Schmid, S. R., and Spies, M. (1988) *J. Biol. Chem.* 263, 16886–16891
- Hermon-Taylor, J., Perrin, J., Grant, D. A. W., Appleyard, A., Bubel, M., and Magge, A. I. (1977) *Gut* 18, 259–265
- Lajda, Z., and Gosarau, R. (1983) *Histochemistry* 78, 251–270
- Hartmann, E., Rapoport, T. A., and Lodish, H. F. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 5786–5790
- Toyoda, S., Lee, P. C., and Lebenthal, E. (1985) *Dig. Dis. Sci.* 30, 1174–1180
- Tomita, M., Furthmayr, H., and Marchesi, V. T. (1978) *Biochemistry* 17, 4756–4770
- Soutar, A. K., and Knight, B. L. (1990) *Br. Med. Bull.* 46, 891–916
- Hunziker, W., Spies, M., Semenza, G., and Lodish, H. F. (1986) *Cell* 46, 227–234
- Watt, V. M., and Yip, C. C. (1989) *J. Biol. Chem.* 264, 5480–5487
- Hamilton, K. K., Zhao, J., and Sims, P. J. (1993) *J. Biol. Chem.* 268, 3632–3638
- Leytus, S. P., Chung, D. W., Kisiel, W., Kurachi, K., and Davie, E. W. (1984) *Proc. Natl. Acad. Sci. U. S. A.* 81, 3699–3702
- Stroud, R. M., Kay, L. M., and Dickerson, R. E. (1974) *J. Mol. Biol.* 83, 185–208
- Baratti, J., and Maroux, S. (1976) *Biochim. Biophys. Acta* 452, 488–496
- Fujikawa, K., Chung, D. W., Hendrickson, L. E., and Davie, E. W. (1986) *Biochemistry* 25, 2417–2424
- Vanderlicke, P., Croik, C. S., Nadel, J. A., and Caughey, G. H. (1989) *Biochemistry* 28, 4148–4155
- Mikeš, O., Holeyšovský, V., Tomášek, V., and Sorm, F. (1966) *Biochem. Biophys. Res. Commun.* 24, 348–352
- Hartley, B. S. (1964) *Nature* 201, 1284–1287
- Meloun, B., Klueh, I., Kostka, V., Motavek, L., Prusik, Z., Vaněček, J., Keil, B., and Sorm, F. (1966) *Biochim. Biophys. Acta* 130, 543–546
- Hartley, B. S., and Kauffmann, D. L. (1966) *Biochem. J.* 101, 229–231
- Blow, D. M., Birktoft, J. J., and Hartley, D. S. (1969) *Nature* 221, 337–340
- Kawashima, I., Tani, T., Shimoda, K., and Takiguchi, Y. (1987) *DNA (N.Y.)* 6, 163–172
- DiScipio, R. G., Chakravarti, D. N., Muller-Eberhard, H. J., and Fey, G. H. (1988) *J. Biol. Chem.* 263, 549–560
- Rao, A. G., Howard, O. M. Z., Ng, S. C., Whitehead, A. S., Colten, H. R., and Sodes, J. M. (1987) *Biochemistry* 26, 3556–3564
- Howard, O. M. Z., Rao, A. G., and Sodes, J. M. (1987) *Biochemistry* 26, 3565–3570
- Herr, J., Hamann, U., Rognes, S., Myklebust, O., Gausepohl, H., and Stanley, K. K. (1988) *EMBO J.* 7, 4119–4127
- Murdoch, A. D., Dodge, G. R., Cohen, L., Tuan, R. S., and Iozzo, R. V. (1992) *J. Biol. Chem.* 267, 8544–8557
- Raychowdhury, R., Niles, J. L., McCluskey, R. T., and Smith, J. A. (1989) *Science* 244, 1163–1165
- Wozney, J. M., Rosen, V., Celeste, A. J., Mitsock, L. M., Whittam, M. J., Krutz, R. W., Hewick, R. M., and Wang, E. A. (1988) *Science* 242, 1528–1534
- Barger, A., and Schechter, I. (1970) *Philos. Trans. R. Soc. Lond. B* 267, 249–264
- von Heijne, G., and Gavel, Y. (1988) *Eur. J. Biochem.* 174, 671–678
- Komooka, H., and Uneyama, H. (1991) *Abstracts of the 14th Symposium on Chemical Information and Computer Science, Kawaguchi*, pp. 71–73. Chemical Society of Japan, Tokyo

Exhibit 23

Perspectives in Bioconjugate Chemistry

EDITED BY
Claude F. Meares
University of California



American Chemical Society, Washington, DC 1993



Library of Congress Cataloging-in-Publication Data

Perspectives in bioconjugate chemistry / edited by Claude F. Meares.

p. cm.

Contains a collection of articles previously published in the journal:
Bioconjugate chemistry.

Includes bibliographical references and index.


ISBN 0-8412-2672-5

1. Bioconjugates.

I. Meares, Claude F., 1946- . II. American Chemical Society.

QP517.B49P47 1993
574.19'2—dc20

93-15385
CIP

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984. 

Copyright © 1993

American Chemical Society

All Rights Reserved. The appearance of the code at the bottom of the first page of each chapter in this volume indicates the copyright owner's consent that reprographic copies of the chapter may be made for personal or internal use or for the personal or internal use of specific clients. This consent is given on the condition, however, that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc., 27 Congress Street, Salem, MA 01970, for copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to copying or transmission by any means—graphic or electronic—for any other purpose, such as for general distribution, for advertising or promotional purposes, for creating a new collective work, for resale, or for information storage and retrieval systems. The copying fee for each chapter is indicated in the code at the bottom of the first page of the chapter.

The citation of trade names and/or names of manufacturers in this publication is not to be construed as an endorsement or as approval by ACS of the commercial products or services referenced herein; nor should the mere reference herein to any drawing, specification, chemical process, or other data be regarded as a license or as a conveyance of any right or permission to the holder, reader, or any other person or corporation, to manufacture, reproduce, use, or sell any patented invention or copyrighted work that may in any way be related thereto. Registered names, trademarks, etc., used in this publication, even without specific indication thereof, are not to be considered unprotected by law.

PRINTED IN THE UNITED STATES OF AMERICA

Chemical Modifications of Proteins: History and Applications

Gary E. Means[†] and Robert E. Feeney[‡]

Department of Biochemistry, The Ohio State University, Columbus, OH 43210, and Department of Food Science and Technology, University of California, Davis, CA 95616

Reprinted from *Bioconjugate Chemistry*, Vol. 1, No. 1, January/February, 1990

With roots in ancient formulations, methods for the chemical derivatization of proteins continue to expand and develop. The creation of this new journal dealing exclusively with bioconjugate chemistry was barely conceivable just a few years ago. An explosion of interest in the subject during the last decade is, however, easily seen. The tremendous growth in both the number of publications and in the number of research groups involved in these kinds of studies has been promoted by both practical interests related, for example, in some cases to possible pharmacological or medical diagnostic applications and by interest in questions of fundamental biochemical structure and function.

Greatly improved understanding of established reagents and procedures and the development of many new, and more sophisticated, reagents and procedures have been facilitated by advances in the ancillary fields of organic chemistry, X-ray crystallography, and molecular biology. Whereas protein modification in the past often involved the same reagents and reactions commonly used in the organic chemistry of that time (i.e., acetylation, iodination, deamination, reaction with formaldehyde, etc.), those in most common use today have, by and large, been developed to meet the varied but relatively specific needs of the protein chemist. A large number of specialized reagents have been described: affinity labels, photoaffinity labels and other specifically designed site-directed reagents (1, 2), group-selective reagents which react exclusively (or at least predominantly) with one particular type of amino acid side chain (see below, especially Table II), and others that react relatively nonspecifically with a number of different side chains (3).

Reagents have been designed to preserve electrostatic charge (4, 5), to alter electrostatic charge (6), and to increase hydrophobicity (7, 8). Reagents and procedures have been developed to decrease immunogenicity (9, 10), to increase and decrease susceptibility to proteolysis (11–13), to increase UV or visible absorbancy (14), to introduce flu-

orescent labels (15, 16), spin labels (17), radiolabels (18–20), various metal ions (21), magnetic microspheres (22, 23), and electron-dense substituents (24), to increase the content of certain low-abundance nonradioactive isotopes (25), and to attach several different types of carbohydrate moieties (26–29), biotin (30), and a number of other biospecific recognition groups (i.e., avidin, streptavidin, antibodies, protein A, protein G, lectins, and others (31)). Procedures also have been developed to effect the cleavage of peptide chains (32, 33); to modify enzyme specificity (34); to modify the terminal hydroxyls of galactosyl residues in glycoproteins (35); to introduce intramolecular and intermolecular cross-links, both to couple already associated species (36, 37); and to join various proteins, which might or might not otherwise associate, in order to combine the properties of both into a single molecule, e.g., to make protein-protein conjugates (38, 39), enzyme-linked antibodies (40, 41), immunotoxins (42, 43), and drug-protein conjugates (44). A large number of reagents that have been developed to serve these and a variety of other purposes are commercially available.

EARLY DEVELOPMENTS

The chemistry of proteins had its origin in the chemistry of the amino acids and only later concerned the amino acid side chains of intact proteins. For practical purposes, a variety of procedures for protein modification had been developed and used many years prior to any significant interest in or understanding of protein chemistry. For example, the use of formaldehyde and other agents in the tanning industry was apparently formulated entirely on the basis of empirical observations, without any real understanding of the reactions or of the chemical nature of the materials involved. Similar procedures were also employed successfully to convert a number of protein toxins, usually of bacterial origin, into toxoids, which retain some of the original antigenic determinants but are no longer toxic. Inoculations of toxoids are still widely employed to confer immunity against a number of serious bacterial diseases. Although still widely

[†] The Ohio State University.

[‡] University of California.

used, there is not much known about the manner by which formaldehyde converts toxins into toxoids.

Interest in quantitative determinations of proteins and their various constituent amino acids was a major impetus for many early studies of chemical modification. While a significant number of proteins had been crystallized by the 1920s, analytical values for individual amino acids were still quite poor well into the 1940s. Analytical data had, for example, revealed only one sulfur-containing amino acid, cystine, in naturally occurring proteins prior to the discovery of methionine in 1922. Threonine was not discovered until 3 years later.

Most of the procedures available at that time for the determination of individual amino acids were, of course, supplanted by the development of the far more convenient cation-exchanger amino acid analyzer in the 1950s. Slightly altered forms of some of those procedures, however, still find use today. Variations of the Van Slyke procedure for determining protein nitrogen, for example, are still sometimes useful for bringing about the selective deamination of proteins. Sodium nitroprusside, which was once used for spectrophotometric determinations of cysteine, also appears to be useful for the selective modification of protein thiol groups. Some much more recently developed procedures for protein modification, on the other hand, have been shown to be useful for analytical determinations of certain amino acids in proteins. The use of water-soluble carbodiimides and certain nucleophiles to determine amounts of glutamine and asparagine, and of 2-hydroxy-5-nitrobenzyl bromide to determine tryptophan contents of proteins are possibly of special interest since the acid lability of those amino acids makes their determinations difficult by conventional amino acid analysis (45, 46). The use of TNBS¹ for the determination of amino groups (47) and DTNB for the determination of thiol groups (48) in intact proteins have also achieved special status as a result of their widespread use for such purposes.

By the end of World War II, interest had turned to determining particular amino acid residues necessary for the biological activities of proteins. That a particular amino acid residue in the active site of an enzyme might be identified on the basis of its reaction with selective chemical reagents was an idea developed during this period. Those interests and further careful scrutiny of the available methodology led to the publication of two important reviews of protein modification in 1947 (49, 50). The report of Balls and Jansen (51) showing that the inactivation of several proteases by diisopropyl fluorophosphate resulted from its reaction with a specific serine residue in each case was another milestone of this period.

Some of the earliest attempts to use chemical modification procedures to identify particular amino acid residues required for the biological activity of a protein were conducted in the laboratory of Heinz Fraenkel-Conrat (52-54). A few of those procedures are still used, with little change, to this day. However, these earlier studies were seriously hampered by the absence of sensitive and accurate procedures to determine the number and type(s) of amino acid residues undergoing modification and by the absence of effective micro and semimicro procedures to separate, purify, and characterize products. The studies of that period, nevertheless, provided important descriptions of procedures for use by other investigators and

served as important steps to the later development of improved procedures.

Quantitative data on the extent of modification became more attainable with the increased availability of radioactively labeled reagents during the 1960s. Greater access to automated amino acid analyzers (55) and the development of effective ion-exchange and gel exclusion chromatography media at about the same time also facilitated the characterization of modified proteins, which led to a better understanding of many modification reagents and procedures. Various forms of micro gel electrophoresis also became commonplace in the same decade, and these greatly enhanced the ability to monitor the effects of modification on relatively small amounts of protein. The advent of an effective procedure for the routine determination of amino acid sequences, first described by Edman in 1956 (56), was also a major milestone. Although often considered routine today, these procedures were developed only after many years of effort and were essential for the characterization of various modification procedures.

SITE-SPECIFIC MODIFICATIONS

In 1962, Wofsey and co-workers (57) described a selective reaction of the *p*-arsonylbenzenediazonium ion with the antigen-combining site of a rabbit anti-*p*-azobenzenearsonate antibody. This demonstration of affinity labeling was followed in about 1 year by the description of a highly selective reaction between chymotrypsin and a reactive substratelike compound, TPCK (58). The latter was shown to effect the modification of a particular histidine residue of chymotrypsin with the complete elimination of its catalytic activity. The selectivity of these and other affinity labels results from their resemblance to a substrate or ligand. Their strong affinity for a particular site concentrates a reactive group, like the chloromethyl ketone moiety of TPCK, at a specific site, where its reaction with a nearby amino acid side chain is promoted by mutual proximity. Subsequent to these reports, a very large number of affinity labeling reagents have been described. Affinity labeling is now one of the most important methods for identifying amino acid residues in enzyme active sites. Table I describes some of the most commonly used types of affinity labeling reagents and summarizes a few of their salient properties.

SIDE CHAIN SELECTIVE MODIFICATIONS

The use of the side chain selective reagents (i.e., those which react, under certain specified conditions, with a single or, at least, a limited number of side-chain groups in a fairly predictable manner) is, however, a simpler approach. At least for initial screening, it is still widely used to identify amino acid chains required for biological activity. Table II contains a list of some of the most commonly used and, in the authors' opinions, most useful group-selective reagents and brief descriptions of some of their important properties and applications.

The retention of biological activity after treatment with one of those reagents is usually good a priori evidence that the modified amino acid side chains are not required for that particular activity. Under appropriate conditions, each reagent normally reacts only with the indicated target side chain(s). Depending on the protein, the reagent, and the particular conditions, however, complete modification of all such side chains is not always obtained. In most cases, the extent of reaction can be determined by either direct spectrophotometric measurements, amino acid analyses, or the use of radioactive

¹ Abbreviations are as follows: trinitrobenzenesulfonic acid, TNBS; 5,5'-dithiobis(2-nitrobenzoic acid), DTNB; tosylphenylalanine chloromethyl ketone, TPCK; dithiothreitol, DTT; 1-ethyl-3-[3-(dimethylamino)propyl]carbodiimide, EDC.

Table I. Major Types of Affinity Labels

type	examples	target enzymes	reaction characteristics	refs cited
α -halocarbonyl $\text{R-COCH}_2\text{X}$	TPCK	chymotrypsin	addition to nucleophilic groups, especially His and Cys(SH), also COO-	58
	3-bromo-2-ketoglutarate chloroacetol sulfate	isocitrate dehydrogenase		59
epoxide $\text{R}-\text{CH}-\text{CH}_2$ $\quad \quad \quad \backslash \quad /$ $\quad \quad \quad \text{O}$	1,2-anhydromannitol 6-phosphate	triose phosphate isomerase	addition to various nucleophilic groups, COO-, Cys(SH)	60 61
	glycidol phosphate	triose phosphate isomerase, enolase		62
sulfonyl fluoride $\text{R-SO}_2\text{F}$	5'-[(fluorosulfonyl)benzoyl]adenosine	glutamine synthetase, etc.	addition to various nucleophilic groups, Cys(SH), Lys, His, etc.	63
aldehyde R-CHO	2',3'-dialdehyde-ATP	pyruvate carboxylase adenylate cyclase, etc.	synthesized by periodate oxidation of ATP, addition to amino groups especially in the presence of NaBH_4 , dialdehyde derivatives of other nucleotides and nucleosides may be employed similarly	64, 65
	pyridoxal phosphate	glycogen phosphorylase, glutamine synthetase, DNA polymerase, etc.	reaction with Lys in PLP and phosphate binding sites; irreversible, in the presence of NaBH_4 or NaBH_3CN	66-68
azido RN_3 (photoaffinity labels)	8-azido-ATP	F1-ATPase	requires UV irradiation; by addition to nucleophiles and double bonds, insertion into C-H and O-H bonds, and other reactions	69
	5-azido-UDP	UDP-glucose, pyrophosphorylase		70

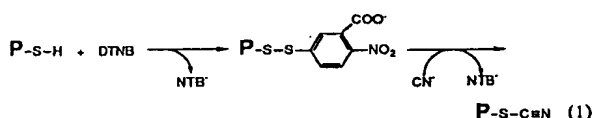
Table II. Useful Side Chain Modification Reagents*

side chain or group	reagent or procedure	optimum reaction pH, side chain selectivity, and other comments	refs cited
amino (Lys + α)	amidation (ethyl acetimidate)	pH ~9, no other side chains react, positive charge maintained, other imido esters are available, extent of modification may be determined with TNBS	4, 71
	reductive alkylation (formaldehyde + NaBH_4 or NaBH_3CN)	pH ~9 with NaBH_4 , pH ~7 with NaBH_3CN ; reaction is much slower under the latter conditions; no other side chains react; positive charge maintained; other aldehydes and reducing agents may be used; extent of modification may be determined by amino acid analysis, the incorporation of radiolabel, or with TNBS	5, 25
	acylation (acetic anhydride)	pH ~8 and above, Tyr residues also modified, elimination of positive charge, extent of modification may be determined with TNBS	72
	(succinic anhydride)	same as above, Tyr residues undergo slow deacylation above pH ~5, replaces positive charges with negative charges	73
	trinitrobenzenesulfonate	pH ~8 and above, also reacts slowly with thiol groups, eliminates positive charge and introduces large hydrophobic substituent, extent of reaction may be determined spectrophotometrically	47, 74
carboxyl (Asp + Glu)	water-soluble carbodiimide + nucleophile (EDC + glycine ethyl ester)	pH ~4.5-5, some side reactions with Tyr and thiol groups, other carbodiimides are available, many other nucleophiles (amines) may be used to either maintain or alter the charge, extent of reaction may be determined by amino acid analysis or from incorporation of radiolabel	45, 75
guanidino (Arg)	dicarbonyls [2,3-butanedione, phenylglyoxal, and (<i>p</i> -hydroxyphenyl)glyoxal]	pH ~7 or higher, reaction promoted by borate buffer, no major side reactions; partially reversible upon dialysis, eliminates positive charge, extent of reaction can be determined from incorporation of radiolabel or by amino acid analysis, other dicarbonyl compounds can also be used (i.e., cyclohexanedione, glyoxal, etc.)	76-79
imidazole (His)	diethyl pyrocarbonate (ethoxyformic anhydride)	pH ~4-5, side reactions with Lys kept to minimum by low pH, extent of modification may be determined by spectrophotometric measurement, reversed in the presence of NH_4OH	80, 81
indole (Trp)	<i>N</i> -bromosuccinimide	usually pH ~4 or lower, higher pH values can be used; thiol groups are rapidly oxidized; Tyr and His react more slowly; extent of modification may be determined spectrophotometrically or by amino acid analysis	82
	2-hydroxy-5-nitrobenzyl bromide	pH <7.5, slight reaction with thiols, strong visible absorbance, can be used to determine the extent of reaction	83, 84
phenol (Tyr)	iodination (I_2^- , chloramine T + I^- , ICl , lactoperoxidase + I^- , and H_2O_2)	pH ~8 or higher, many different procedures and reagents, His also reacts but usually to a lesser extent, thiol groups are rapidly oxidized, both mono and diiodo derivatives are formed, the extent of reaction can be estimated spectrophotometrically or by amino acid analysis, widely used for radiolabeling of proteins	18, 85, 86
	tetranitromethane	pH ~8 or slightly higher, thiol groups are also rapidly oxidized, some nitration of Trp, extent of reaction may be determined spectrophotometrically or by amino acid analysis	87
thiol (Cys-SH)	carboxymethylation (iodo- and bromoacetate and iodo- and bromoacetamide)	pH ~7 or higher; no effect on other residues under appropriate conditions; Lys, His, Tyr and Met react slowly with excess reagent and long reaction times; extent of reaction may be determined with DTNB, by the incorporation of radiolabel, or by amino acid analysis	88, 89
	<i>N</i> -ethylmaleimide	pH ~8 or higher, reaction with Lys and His are much slower at pH 7 and usually of no importance, the extent of reaction may be determined from incorporation of radiolabel or by amino acid analysis	90, 91
	5,5'-dithiobis(2-nitrobenzoic acid) (Ellman's reagent)	pH ~7 or higher, no other side chains react, reversible in presence of excess low MW thiol, the extent of modification can be determined spectrophotometrically	48, 92
thioether (Met)	oxidation (H_2O_2)	pH ~1 and higher, thiol groups also react very rapidly, reversed by treatment with low MW thiols, extent of modification may be determined by amino acid analysis after alkaline hydrolysis or by carboxymethylation followed acid hydrolysis	93

* Many useful reagents have not been included due to space limitations. Descriptions of reaction conditions, outcomes and literature citations are also brief and incomplete for the same reason. More complete information is available in the references and other sources cited elsewhere in this review.

reagents. Indirect determinations can also be obtained from the number of unreacted amino acid residues, as determined either spectrophotometrically (e.g., amino groups by TNBS (47) or thiol groups by DTNB (48)) or by amino acid analysis. The extent of reaction can, of course, almost always be increased by the use of more vigorous reaction conditions, e.g., longer reaction times, larger excesses of reagent, and the presence of urea or other denaturing agents. Using more severe conditions, however, is usually accompanied by some decrease in side-chain selectivity, greater risk of conformational change, and, sometimes, other disadvantages. Reaction with other than target side chains may be of little importance when activities are not affected.

A major loss of biological activity upon such treatment is often taken as evidence for the essentiality of the group modified. But this interpretation must be made with somewhat less conviction, owing to the possibility of unrecognized conformational changes or other subtle effects that may always accompany the modification of a protein. The latter are obviously of less concern when fewer side chains are modified and for those modifications that effect the least change in the size and character of side chains. Luckily, a reasonable number of reagents are available for some of the more important side chains, allowing some discretion as to the nature of the modifications that may be effected. Rat liver glycine methyltransferase, for example, is completely inactivated by reaction with excess DTNB (94). The inactivated enzyme is, however, almost completely reactivated by subsequent treatment with potassium cyanide which, presumably, brings about the replacement of a relatively large and anionic 2-nitro-5-thiobenzoate moiety by a smaller cyano group with no formal charge, as follows:



A carboxymethyl moiety introduced by reaction with iodoacetate is also anionic but intermediate in size and effects only a partial loss of activity. The larger groups thus appear to block or otherwise perturb the active site, although none of the cysteine residues to which they are attached are really essential for catalytic activity.

Similar inactivations have been noted following the addition of large or charged groups to the cysteine residues of many enzymes that are either not inactivated or are only partially inactivated by the addition of smaller groups. 2-Nitro-5-thiocyanatobenzoic acid can be used to effect a direct, single-step addition of cyano moieties to thiol groups (95, 96), although its reactions are not quite as simple as they might initially seem (97). Another reagent, methyl methanethiosulfonate, can be used to attach relatively small, uncharged thiomethyl groups to cysteine residues, usually with comparable results (98).

As a general rule, modifications that have the least effect on side-chain character should have the least effect on protein structure and properties. Modifications of lysine residues that retain their usual cationic charge have, for example, generally been found to have relatively little effect on the biological activities and other properties of many proteins. Complete guanidination of the ϵ -amino groups in tuna heart cytochrome *c* thus has almost no effect on its UV-visible spectrum, its redox potential, or its activity in a standard succinate oxidase assay system

(99). The catalytic activity of papain is also essentially unaffected by complete guanidination (100). Amidination or reductive alkylation of amino groups, both of which also retain the cationic charge, are generally preferred today, however, as both of those reactions take place under milder conditions (4, 5, 25).

SIDE-CHAIN REACTIVITIES

The reactivities of side-chain groups in proteins vary considerably depending on their locations and the influence of nearby residues with which they interact. Under appropriate conditions, differences in reactivity can be used to characterize the environments of such side-chain groups. Kaplan and co-workers (101, 102) and others (103, 104), for example, have developed procedures to determine the relative reactivities of certain types of side chains from the extent of their reaction with trace levels of one of several simple reagents. The intrinsic reactivity and pK_a of each reacting group can be determined by comparing its reaction to that of a simple model compound over a range of pH values.

For identical side-chain groups at different sequence positions, the observed differences in pK_a and reactivity are assumed to reflect differences in local environment. Side chains that experience a change in environment upon the binding of a ligand, complexation with another protein, a change in redox state, or the like can be identified by comparing the extent of their reaction in the two different states. This approach has been used primarily to evaluate the environments of the nucleophilic side chains—amino groups and histidine and tyrosine side chains—in proteins (105, 106).

Different local environments may either suppress or enhance the reactivities of individual side-chain groups. Unusually reactive side chains are usually relatively easy to distinguish from others on the basis of their reactivity and are, in many cases, also those required for biological activity. Rates of inactivation, which may differ from overall rates of modification, can be used in many cases to characterize the reactivity and, sometimes, the number of active site residues (107–109).

In many relatively simple cases, rates of inactivation can be correlated with those for the modification of one or more individual amino acid residues. The catalytic subunit of rabbit muscle cAMP-dependent protein kinase, for example, has only two thiol groups, and undergoes a biphasic reaction with DTNB (110). Its rapid inactivation under those conditions correlates with the initial, rapid phase of modification, which has been shown to reflect the reaction of one thiol group about 17 times faster than the other. In this and other cases where rates of inactivation exceed overall rates of modification, selectively labeled derivatives, modified only at the active site, can often be isolated and characterized (111–113).

Activities remaining at various stages of partial modification can also be used, in some cases, to estimate the number of essential residues according to a procedure first described by Tsou in 1962 (114). The decreased iron-binding capacity of chicken egg white ovotransferrin after partial modification by phenylglyoxal, for example, suggests an arginine residue is required for each of its two bound Fe^{3+} ions (76). In the more complicated case of transketolase, two arginine residues per dimer appear to be required for activity, but one appears to react with phenylglyoxal about 40 times faster than the other (115).

SPECTROSCOPIC AND FLUORESCENT LABELS

A number of important procedures requiring the incorporation of spectroscopic or fluorescent labels have been

developed to characterize certain structural features of proteins. Fluorescence lifetimes and quantum yields of many different fluorescent groups and their sensitivities to quenching by acrylamide, iodide, and other substances can, for example, be used to evaluate environments in the vicinity of residues to which those groups have been attached (15, 116). Fluorescence energy transfer measurements are also widely employed to estimate distances between certain internal, or intrinsic, chromophores and various selectively introduced, extrinsic, fluorescent labels and, in some cases, between selectively introduced, extrinsic, donor-acceptor pairs (117, 118). Iodoacetamidofluorescein, dansyl chloride, and *N*-1-pyrenylmaleimide are three examples from a very large number of fluorescent labels that have been used for such purposes. Most may be considered to be analogues of commonly used group-selective reagents and their reaction characteristics may be predicted accordingly.

An extensive list of such reagents, with brief descriptions of their principal reaction and emission and excitation characteristics, has been presented by Haugland (119). Procedures to attach nitroxide moieties, for example the reaction of 4-(2,2,6,6-tetramethyl-1-oxypiperidin-4-yl)-2-(fluorosulfonyl)benzamide with chymotrypsin, have also been employed to obtain information concerning the protein environment and to detect conformational changes by EPR spectroscopy (17, 120).

CROSS-LINKING AND IMMOBILIZATION

Cross-linking of proteins and their immobilization, either by attachment to an insoluble support or by various other means, have a long and important history. The former is sometimes employed to increase the stability of proteins or of certain conformational relationships in proteins, to couple two or more different proteins (e.g., to join different activities into a single molecule), to identify or characterize the nature and extent of certain protein-protein interactions, and, in other cases, to determine distances between reactive groups in or between protein subunits (36, 37, 121-125). Proteins are sometimes immobilized to facilitate their reuse and their separation from other products and (in some cases) to increase their stability. A large number of different procedures, including physical as well as chemical procedures, have been developed to immobilize proteins, and many reviews, symposia proceedings, and books on this subject are available (126-130).

A large number of different types of cross-linking or, as they are sometimes called, bifunctional reagents have been described. They include so-called zero-length cross-linking agents that bring about the direct formation of covalent bonds between existing amino acid side chain groups. The use of water-soluble carbodiimides to bring about the formation of amide linkages between carboxyl groups of aspartate or glutamate and the ϵ -amino groups of lysine side chains appear to be the most prominent zero-length cross-linking agents (123, 131-133). Disulfide bonds obtained from existing thiol groups would also, presumably, be considered zero-length cross-links (134, 135). Such linkages appear to be formed only when the reacting groups are in close proximity.

Other cross-linking agents may be organized according to the type(s) of reactive groups, their side chain reactivity, their hydrophobicity or hydrophilicity, and the length or distance between the reactive groups; whether the two, or in some cases more (136), reactive groups are the same or different (i.e., "homobifunctional" or "heterobifunctional" reagents), whether the structure con-

necting the reactive groups is readily cleavable, and whether the groups are membrane permeable or impermeable, and according to various other criteria. A list of the most widely used types of cross-linking agents and a few brief comments on some of their significant properties are presented in Table III. A much more extensive list of cross-linking agents has been presented by Ji (125).

The reactivities of cross-linking agents, except for one or two special cases, are very similar to those of the corresponding monofunctional reagents. The initial reaction with a protein is presumably, in most cases, a simple second-order process, not seriously affected by the second reactive group. The latter's reaction, however, is completely dependent on the availability of a second appropriate side chain which, for fast, efficient cross-linking, must be both nearby and in an appropriate orientation. Cross-linking agents with different lengths, different stereochemical configurations (some with little and others with a great deal of conformational flexibility), and with different side-chain specificities have been developed to fulfill different needs. Distances between potentially reactive side chains in the same or different subunits of some oligomeric proteins have, for example, been estimated by comparing rates and yields of cross-link formation with a series of cross-linking agents differing in length, stereochemical configuration, and side-chain reactivity (139, 155, 146).

The importance of side-chain proximity in these reactions is perhaps most evident in the case of cross-linking agents that undergo hydrolysis or some other inactivation process in addition to their cross-linking of proteins. The use of bifunctional imidoesters to characterize oligomeric proteins, for example, is based on the formation of recognizable SDS gel electrophoretic patterns, reflecting the formation of cross-links between adjacent subunits (139, 138). Like the cross-links within a subunit, those between subunits are formed only when two amino groups are in close and appropriate proximity. Cross-links between other than adjacent subunits are largely precluded by the hydrolytic instability of the monofunctional imidoester intermediates. The importance of hydrolytic stability on yields of cross-linked products has been discussed by Staros (37, 156).

Of the 20 or so amino acid side chains normally present in proteins, ϵ -amino groups of lysine residues are usually among the most abundant and most accessible of the potentially reactive groups. A relatively large proportion of the most commonly used cross-linking agents are therefore amino group selective reagents (i.e., imidoesters, *N*-hydroxysuccinimide esters, activated aryl fluorides, etc.). Most of them, however, also undergo fairly rapid hydrolysis in addition to their reaction with amino groups, which, except for cases involving close proximity, seriously limits the yields that may be obtained. Glutaraldehyde, which does not hydrolyze or become otherwise inactivated over long periods of time, is widely used to immobilize enzymes by cross-linking and to stabilize their adsorption to or entrapment in various materials (157, 158). The nature of its reactions with proteins may involve some Schiff base formation but is clearly much more complicated than that and not completely understood (137, 159, 160).

The high reactivities of thiol groups with *N*-ethylmaleimide, iodoacetate, and many related α -halocarbonyl compounds has led to the development of many cross-linking agents containing comparable maleimide and α -halocarbonyl moieties. Under the conditions usually employed for cross-linking, the latter are much more sta-

Table III. Homobifunctional and Heterobifunctional Protein Cross-Linking Agents*

agent	description	refs cited
Homobifunctional		
glutaraldehyde	available as 25% aqueous solution, very effective reaction with amino groups and perhaps other nucleophilic groups, contains polymeric and other unknown materials, the nature of the reaction(s) are not known, slow progressive changes proceed long after the initial irreversible coupling	137
dimethyl suberimidate (DMS)	a water-soluble solid; reacts only with amino groups and does not eliminate their cationic charge; reaction at pH 8 or above (optimal at pH ~9); $t_{1/2} \approx 46$ min at pH 8.5 and 25 °C; ~11-Å span; many related reagents with different spans, some readily cleavable, are available or can be easily synthesized	138, 139
disuccinimidyl suberate (DSS)	a water-insoluble solid; must usually be dissolved in DMSO or other water-miscible organic solvent; reacts with amino groups at pH 7 or above; reaction rates increase with pH; $t_{1/2} \approx 4-5$ h at pH 7; ~11-Å span; many related reagents with different spans; hydrophilic spacer arms, some cleavable and water-soluble; sulfosuccinimide esters are available	140, 141
bismaleimidoethane (BMH)	a water-insoluble solid, must usually be dissolved in DMF or other water-miscible organic liquid; reacts with thiol groups at pH ~6-8; ~16-Å span; many related reagents with different span lengths; more hydrophilic spacer arms and cleavable analogs are available	142, 143
<i>p</i> -phenylenemaleimide	a water-insoluble solid, must usually be dissolved in water-miscible organic solvent, reacts with thiol groups at pH ~6-8; ~12-Å span, ortho and meta isomer are also available, less stable than aliphatic maleimides	144-146
Heterobifunctional		
<i>m</i> -maleimidobenzoic acid <i>N</i> -hydroxysuccinimide ester (MBS)	a water-insoluble solid, must usually be dissolved in water-miscible organic liquid, initial reaction with amino group component at pH ~7-8 followed by coupling with thiol component at pH ~6-8, ~10-Å span, more water soluble sulfosuccinimide ester is also available	147, 148
<i>N</i> -succinimidyl 4-(<i>N</i> -maleimidomethyl)- cyclohexane-1-carboxylate (SMCC)	a water-insoluble solid, must usually be dissolved in water-miscible organic solvent, reaction characteristics very similar to those of MBS, ~12-Å span, more water soluble sulfosuccinimide ester is also available	149, 150
<i>N</i> -succinimidyl 3-(2-pyridyldithio)propionate (SPDP)	a water-insoluble solid, must usually be dissolved in a water-miscible organic solvent, initial reaction with the amino component at pH ~7-8.5 followed by either coupling to thiol component at pH 7 or above or treatment with DTT followed by coupling to maleimidylated protein, ~7-Å span	151, 152
2-iminothiolane ("Traut's reagent")	a water-soluble solid; reacts only with amino groups at pH 7-10 without eliminating their charge; reaction may be followed with DTNB; ~8-Å span; may be coupled directly to MBS-, SMCC- or SPDP-treated proteins	153, 154

* Many more cross-linking agents have been described. Those included appear to be among the most widely used and most important at the present time. Please consult references in the text for additional examples.

ble to hydrolysis than the amino group reagents mentioned above and the yields of cross-linked products are, therefore, usually somewhat less dependent on side chain proximity (161, 162).

A large number of heterobifunctional cross-linking reagents have been developed which usually contain a thiol reactive and an amino group reactive moiety. *N*-Alkyl- or *N*-arylmaleimide and α -halocarbonyl groups are the most common of the former and *N*-hydroxysuccinimide esters appear to be the most common of the latter. To increase aqueous solubility, sodium salts of sulfonated *N*-hydroxysuccinimide esters are also commonly employed (163). In addition to the two reactive groups a variety of different types of connecting structures or spacer arms have been employed. The nature of the spacer arm may, of course, also have important consequences. Longer spacer arms are usually assumed to be more effective for coupling larger proteins or those where the potentially reactive side chains are sterically protected. The conformational flexibility, hydrophilicity or hydrophobicity, and the "cleavability" of the spacer arm are also important considerations. *N*-Alkylmaleimides are also generally more stable than their aryl counterparts (162, 164).

Photoactivatable heterobifunctional cross-linking agents are particularly useful for identifying interacting components in complicated biological systems (165). Wood and O'Dorisio (166), for example, used *N*-succinimidyl 4-azido-2'-nitrophenyl-6-[(4'-azido-2'-nitrophenyl)-amino]hexanoate and two nonphotoactivatable homobifunctional cross-linking agents to identify vasoactive intestinal peptide receptors in human lymphoblasts by their coupling to ¹²⁵I-labeled vasoactive intestinal peptide. A

photoactive derivative of a *N*-formylated chemotactic peptide, prepared by reaction with the last mentioned photoactivatable agent, has also been used to characterize the *N*-formyl peptide receptors of human polymorphonuclear leukocytes (167).

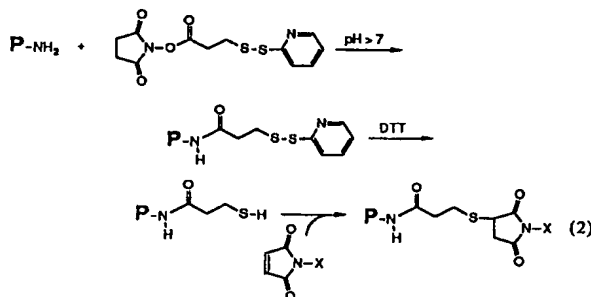
The initial reaction with photoactivatable cross-linking agents is usually conducted in the dark so that the photoreactive group is inert. Cross-linking is then initiated in a subsequent step involving exposure to light. Azido groups which are converted into a highly reactive nitrenes and diazo moieties (i.e., diazoacetyl, diazo ketones, etc.) which give even more reactive carbenes upon photoactivation are the most common photoactivatable groups in use at this time (2, 3). Being so reactive, both react relatively indiscriminately with OH, NH, CH, and C=C moieties in their vicinity and have short half-lives. Their reaction with surrounding solvent usually precludes reaction with groups not in their immediate vicinity and leads to quite low yields. The detection of cross-linked products thus often provides a good record of spatial relationships at the moment of photolysis but the yields are not adequate for most preparative purposes.

Heterobifunctional cross-linking agents are particularly useful for conjugating different proteins. The different side-chain reactivities of the two reactive groups, for example, usually permit the coupling to be carried out in a stepwise manner which allows, in some cases, for partial purification and, if desired, characterization of intermediates prior to the actual conjugation. Due to the hydrolytic instability of the most important groups directed at amino side chains, the first step usually involves addition of the cross-linker to the amino groups of one member of the future hybrid pair (which either has no

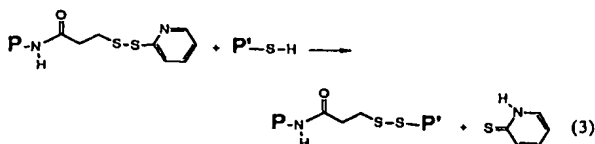
thiol groups or where thiols, if present, are at least temporarily blocked). The removal of unreacted or hydrolyzed reagent and other unwanted substances is usually possible at this stage. The resulting derivative is then directly coupled via the introduced thiol-reactive maleimido or α -halocarbonyl group(s) to the thiol-containing member of the intended hybrid pair.

An artificial antibody-ricin conjugate, for example, has been prepared by treating ricin with *m*-maleimidobenzoyl *N*-hydroxysuccinimide ester and then incubating the resulting *m*-maleimidobenzoyl derivative with a partially reduced monoclonal antibody (148). The formation of unwanted homoprotein conjugates is precluded by such two-step procedures, and purification of the resulting hybrid conjugates by exclusion chromatography is usually rather easy since they should be significantly larger than any of their precursors. Iodoacetyl derivatives of avidin, alkaline phosphatase, and at least four other proteins are commercially available.

Several reagents have been employed to introduce thiol groups into proteins, which may then be employed for conjugation to other proteins or various other materials. *N*-Acetylhomocysteine thiolactone (168), (*S*-acetylthio)succinic anhydride (169), *S*-acetyl *N*-succinimidylthioacetate (170), 2-iminothiolane (153), and *N*-succinimidyl 3-(2-pyridyldithio)propionate (151), for example, can all be used under mildly alkaline conditions to introduce thiol groups into proteins. In the second and third cases, the acetyl moiety must subsequently be removed, usually by treatment with hydroxylamine, to release the thiol group and, in the last case, a small amount of DTT or some other simple thiol must be used to affect a comparable cleavage of the 2-pyridyl disulfide moiety. The resulting thiol groups potentially can be coupled to many different maleimidyl or α -halocarbonyl groups including, for example, those of certain protein-maleimidyl conjugates as follows (171, 150):



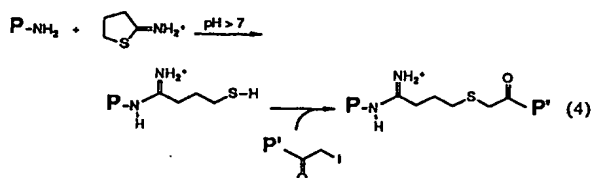
Even more important, probably, is the ability of the latter substituent to undergo direct coupling with the thiol groups of other proteins as follows (152, 172):



Several 2-pyridyl disulfide-protein conjugates are commercially available. The susceptibility of disulfide linkages to cleavage by low molecular weight thiols, however, appears to preclude many applications of such con-

jugates, including most of those involving exposure to physiological conditions.

2-Iminothiolane is probably the most important reagent for introducing thiol groups into proteins. It is quite water soluble, whereas the others really are not, it reacts rapidly with amino groups at pH 7 (or preferably a little above), and it does not require an additional activation step to effect release of the thiol moiety. It alone preserves the cationic charges of the modified amino groups. As with the other reagents used to introduce thiol groups, those introduced via reaction with 2-iminothiolane can be used to effect oxidative coupling to other protein thiols or may react with various maleimidyl or α -halocarbonyl groups, as follows (173, 154):



CONCLUSION

Space and time limitations have precluded the discussion of many important related subjects. We had hoped, in particular, to discuss the radiolabeling of proteins. Biotinylation also deserves serious discussion. We apologize to the many authors whose works we have failed to cite and particularly to those whose results we may have misinterpreted or misrepresented. We would also like to call the readers' attention to a number of reviews and books on this subject, where more complete information can be obtained (174-183).

ACKNOWLEDGMENT

Financial support to GEM was received from Solar Energy Research Institute and to REF from U.S. National Institutes of Health Grant GM23817. The assistance of Shirley Miller in preparing the manuscript is greatly appreciated.

LITERATURE CITED

- (1) Colman, R. F. (1983) Affinity labeling of purine nucleotide sites of proteins. *Annu. Rev. Biochem.* 52, 67-91.
- (2) Bayley, H. (1983) *Photogenerated Reagents in Biochemistry and Molecular Biology* Elsevier, New York.
- (3) Knowles, J. R. (1972) Photogenerated reagents for biological receptor-site labels. *Acc. Chem. Res.* 5, 155-160.
- (4) Hunter, M. J. and Ludwig, M. L. (1962) The reaction of imidoesters with proteins and related small molecules. *J. Am. Chem. Soc.* 84, 3491-3504.
- (5) Means, G. E. and Feeney, R. E. (1968) Reductive alkylation of amino groups in proteins. *Biochemistry* 7, 1366-1371.
- (6) Goldstein, L., Leven, Y., and Katchalski, E. (1964) A water-insoluble polyanionic derivative of trypsin. II. Effect of the polyelectrolyte carrier on the kinetic behavior of bound trypsin. *Biochemistry* 3, 1913-1919.
- (7) Nishikawa, A. H., Morita, R. Y., and Becker, R. R. (1968) Effects of the solvent medium on polyvalylribonuclease aggregation. *Biochemistry* 7, 1506-1513.
- (8) Ampon, K. and Means, G. E. (1988) Immobilization of proteins on organic polymer beads. *Biotechnol. Bioeng.* 32, 689-697.
- (9) Abuchowski, A., van Es, T., Palczuk, N. C., and Davis, F. F. (1977) Alteration of immunological properties of bovine serum albumin by covalent attachment of polyethylene glycol. *J. Biol. Chem.* 252, 3578-3581.

- (10) Veronese, F. M., Largalolli, R., Boccu, E., Bengassi, C. A., and Schiavon, O. (1985) Surface modification of proteins. Activation of monomethoxy-polyethylene glycols by phenylchloroformates and modification of ribonuclease and superoxide dismutase. *Appl. Biochem. Biotechnol.* 11, 141-152.
- (11) Raftery, M. A. and Cole, R. D. (1966) On the aminoethylation of proteins. *J. Biol. Chem.* 241, 3457-3461.
- (12) Dixon, H. B. F. and Perham, R. N. (1968) Reversible blocking of amino groups with citraconic anhydride. *Biochem. J.* 109, 312-314.
- (13) Rice, R. H., Means, G. E., and Brown, W. D. (1977) Stabilization of bovine trypsin by reductive methylation. *Biochem. Biophys. Acta* 492, 316-321.
- (14) Parkinson, D. and Redshaw, J. D. (1984) Visible labeling of proteins for polyacrylamide gel electrophoresis with dabyl chloride. *Anal. Biochem.* 141, 121-136.
- (15) Hudson, E. N. and Weber, G. (1973) Synthesis and characterization of two fluorescence sulfhydryl reagents. *Biochemistry* 12, 4154-4161.
- (16) Weltman, J. K., Szaro, R. P., Frackelton, A. R., Dowben, R. M., Bunting, J. R., and Cathow, R. E. (1973) *N*-(3-Pyrene)maleimide: a long lifetime fluorescent sulfhydryl reagent. *J. Biol. Chem.* 248, 3173-3177.
- (17) Berliner, L. J. and Wong, S. S. (1974) Spin-labeled sulfonyl fluorides as active site probes of protease structure. *J. Biol. Chem.* 249, 1668-1682.
- (18) Hunter, W. M. and Greenwood, F. C. (1962) Preparation of I-131 labelled human growth hormone of high specific activity. *Nature* 194, 492-496.
- (19) Rice, R. H. and Means, G. E. (1971) Radioactive labeling of protein in vitro. *J. Biol. Chem.* 246, 831-832.
- (20) Bolton, A. E. and Hunter, W. M. (1973) The labelling of proteins to high specific radioactivities by conjugation to a ¹²⁵I-containing acylating agent. *Biochem. J.* 133, 529-539.
- (21) Meares, C. F., McCall, M. J., Deshpande, S. V., DeNardo, S. J., and Goodwin, D. A. (1988) Chelate radiochemistry: Cleavable linkers lead to altered levels of radioactivity in the liver. *Int. J. Cancer* 2, 99-102.
- (22) Langer, R. and Brown, E. (1985) Controlled release and magnetically modulated release systems for macromolecules. *Methods Enzymol.* 112, 399-422.
- (23) Senyei, D. and Widder, K. J. (1985) Biophysical drug targeting: Magnetically responsive albumin microspheres. *Methods Enzymol.* 112, 56-67.
- (24) Petsko, G. A. (1985) Preparation of heavy-atom derivatives. *Methods Enzymol.* 114, 147-156.
- (25) Jentoft, N. and Dearborn, D. G. (1979) Labeling of proteins by reductive methylation using sodium cyanoborohydride. *J. Biol. Chem.* 254, 4359-4365.
- (26) Maekawa, K. and Liener, I. E. (1960) Properties of the glucosylamidyl derivative of trypsin. *Arch. Biochem. Biophys.* 91, 101-107.
- (27) Lee, H. S., Sen, L. C., Clifford, A. J., Whitaker, J. R., and Feeney, R. E. (1979) Preparation and nutritional properties of caseins covalently modified with sugars. Reductive alkylation of lysines with glucose, fructose or lactose. *J. Agric. Food Chem.* 27, 1094-1098.
- (28) Chen, V. J. and Wold, F. (1984) Neoglycoproteins: preparation of noncovalent glycoproteins through high-affinity protein-(glycosyl) ligand complexes. *Biochemistry* 23, 3306-3311.
- (29) Wong, W. S. D., Kristjansson, M. M., Osuga, D. T., and Feeney, R. E. (1985) 1-Deoxyglycitulation of protein amino groups and their regeneration by periodate oxidation. *Int. J. Peptide Protein Res.* 26, 55-62.
- (30) Hofmann, K., Titus, G., Montibeller, J. A., and Finn, F. M. (1982) Avidin binding of carboxyl-substituted biotin analogues. *Biochemistry* 21, 978-984.
- (31) Wilchek, E. A. and Bayer, E. A. (1988) The avidin-biotin complex in bioanalytical applications. *Anal. Biochem.* 171, 1-32.
- (32) Gross, E. (1967) The cyanogen bromide reaction. *Methods Enzymol.* 11, 238-255.
- (33) Mahoney, W. C. and Hermodson, M. A. (1979) High-yield cleavage of tryptophan peptide bonds by *o*-iodosobenzoic acid. *Biochemistry* 18, 3810-3814.
- (34) Kaiser, E. T., Lawrence, D. S., and Rokita, S. E. (1985) The chemical modification of enzymatic specificity. *Annu. Rev. Biochem.* 54, 565-595.
- (35) Osuga, D. T., Feather, M. S., Shah, M. J., and Feeney, R. E. (1989) Modification of galactose and *N*-acetylgalactosamine residues by oxidation of C-6 hydroxyls to the aldehydes followed by reductive amination: Model systems and antifreeze glycoproteins. *J. Protein Chem.* 8, 519-528.
- (36) Han, K.-K., Richard, C., and Delacorte, A. (1984) Chemical cross-links of proteins by using bifunctional reagents. *Int. J. Biochem.* 16, 129-145.
- (37) Staros, J. V. (1988) Membrane-impermeant cross-linking reagents: Probes of structure and dynamics of membrane proteins. *Acc. Chem. Res.* 21, 435-441.
- (38) Poznansky, M. J. (1986) Tailoring Proteins for More Effective Use as Therapeutic Agents. In *Protein Tailoring for Food and Medical Uses* (R. E. Feeney and J. R. Whitaker, Eds.) pp 317-337, Marcel Dekker, New York.
- (39) Poznansky, M. (1988) Soluble enzyme conjugates: New possibilities for enzyme replacement therapy. *Methods Enzymol.* 137, 566-574.
- (40) Bode, C., Runge, M. S., Newell, J. B., Matsueda, G. R., and Haber, E. (1987) Characterization of an antibody-urokinase conjugate, a plasminogen activator targeted to fibrin. *J. Biol. Chem.* 262, 10819-10823.
- (41) Beyzavi, K., Hampton, S., Kwasowski, P., Fickling, S., Marks, V., and Clift, R. (1987) Comparison of horseradish peroxidase and alkaline phosphatase-labelled antibodies in enzyme immunoassays. *Ann. Clin. Biochem.* 24, 145-162.
- (42) Cumber, A. J., Forrester, J. A., Foxwell, B. M. J., Ross, W. C. J., and Thorpe, P. E. (1985) Preparation of antibody-toxin conjugates. *Methods Enzymol.* 112, 207-225.
- (43) Faulstich, H. and Fiume, L. (1985) Protein conjugates of fungal toxins. *Methods Enzymol.* 112, 225-237.
- (44) Urdahl, D. L. and Hakomori, S. (1980) Tumor-associated ganglio-*N*-triosylceramide target for antibody dependent, avidin mediated drug killing of tumor cells. *J. Biol. Chem.* 255, 10509-10516.
- (45) Hoare, D. G. and Koshland, D. E. (1967) A method for the quantitative modification and estimation of carboxylic acid groups in proteins. *J. Biol. Chem.* 242, 2447-2453.
- (46) Barman, T. E. and Koshland, D. E. (1967) A colorimetric procedure for the quantitative determination of tryptophan residues in protein. *J. Biol. Chem.* 242, 5771-5776.
- (47) Fields, R. (1972) The rapid determination of amino groups with TNBS. *Methods Enzymol.* 25, 464-468.
- (48) Ellman, G. L. (1959) Tissue sulfhydryl groups. *Arch. Biochem. Biophys.* 82, 70-77.
- (49) Olcott, H. S. and Fraenkel-Conrat, H. (1947) Specific group reagents for proteins. *Chem. Rev.* 41, 151-197.
- (50) Herriott, R. M. (1947) Reactions of native proteins with chemical reagents. *Adv. Protein Chem.* 3, 161-225.
- (51) Balls, A. K. and Jansen, E. F. (1952) Stoichiometric inhibition of chymotrypsin. *Adv. Enzymol.* 13, 321-343.
- (52) Fraenkel-Conrat, H., Bean, R. S., and Lineweaver, H. (1949) Essential groups for the interaction of ovomucoid (egg white trypsin inhibitor) and trypsin, and for tryptic activity. *J. Biol. Chem.* 177, 385-403.
- (53) Fraenkel-Conrat, H. and Olcott, H. S. (1948) The reaction of formaldehyde with proteins. V. Cross-linking between amino and primary amide or guanidyl groups. *J. Am. Chem. Soc.* 70, 2673-2684.
- (54) Fraenkel-Conrat, H. and Feeney, R. E. (1950) The metal-binding activity of conalbumin. *Arch. Biochem.* 29, 101-113.
- (55) Moore, S. and Stein, W. H. (1963) Chromatographic determination of amino acids by the use of automatic recording equipment. *Methods Enzymol.* 6, 819-831.
- (56) Edman, P. and Begg, G. (1967) A protein sequenator. *Eur. J. Biochem.* 1, 80-91.
- (57) Wofsy, L., Metzger, H., and Singer, S. J. (1962) Affinity labeling—A general method for labeling the active sites of

- antibody and enzyme molecules. *Biochemistry* 1, 1031-1039.
- (58) Schoellmann, G. and Shaw, E. (1963) Direct evidence for the presence of histidine in the active center of chymotrypsin. *Biochemistry* 2, 252-255.
 - (59) Ehrlich, R. S. and Colman, R. F. (1987) Characterization of an active site peptide modified by the substrate analogue 3-bromo-2-ketoglutarate on a single chain of dimeric NADP⁺ dependent isocitrate dehydrogenase. *J. Biol. Chem.* 262, 12614-12619.
 - (60) Hartman, F. C., LaMuraglia, C. M., Tomozawa, Y., and Wolfenden, R. (1975) The influence of pH on the interaction of inhibitors with triosephosphate isomerase and determination of the pK_a of the active-site carboxyl group. *Biochemistry* 14, 5274-5291.
 - (61) O'Connell, E. L. and Rose, I. H. (1973) Affinity labeling of phosphoglucose isomerase by 1,2-anhydroisitol-6-phosphates. *J. Biol. Chem.* 248, 2225-2231.
 - (62) Schray, K. J., O'Connell, E. L., and Rose, I. A. (1973) Inactivation of muscle triose phosphate isomerase by D- and L-glycidolphosphate. *J. Biol. Chem.* 248, 2214-2218.
 - (63) Pinkofsky, H. D., Ginsburg, A., Reardon, J., and Heinrikson, R. L. (1984) Lysyl residue 47 is near the subunit ATP-binding site of glutamine synthetase from *Escherichia coli*. *J. Biol. Chem.* 259, 9616-9622.
 - (64) Easterbrook-Smith, B., Wallace, J. C., and Keech, D. B. (1976) Pyruvate carboxylase: Affinity labelling of the magnesium adenosine triphosphate binding site. *Eur. J. Biochem.* 62, 125-130.
 - (65) Wescott, K. R., Olwin, B. B., and Storm, D. P. (1980) Inhibition of adenylate cyclase by the 2'-3'-dialdehyde of adenosine triphosphate. *J. Biol. Chem.* 255, 8767-8776.
 - (66) Fischer, E. H., Kent, E. B., Snyder, E. R., and Krebs, E. G. (1958) The reaction of sodium borohydride with muscle phosphorylase. *J. Am. Chem. Soc.* 80, 2906-2907.
 - (67) DiIanni, C. L. and Villafranca, J. J. (1989) Identification of amino acid residues modified by pyridoxal 5'-phosphate in *Escherichia coli* glutamine synthetase. *J. Biol. Chem.* 264, 8686-8691.
 - (68) Basu, A., Kedar, P., Wilson, S., and Modek, M. J. (1989) Active-site modification of mammalian DNA polymerase β with pyridoxal 5'-phosphate. Mechanism of inhibition and identification of lysine 71 in the deoxynucleoside triphosphate binding pocket. *Biochemistry* 28, 6305-6309.
 - (69) Hollemans, M., Runswick, M. J., Fearnley, I. M., and Walker, J. E. (1983) The sites of labeling of the β -Subunit of bovine mitochondrial F1-ATPase with 8-azido-ATP. *J. Biol. Chem.* 258, 9307-9313.
 - (70) Drake, R. D., Evans, R. K., Wolf, M. J., Haley, B. E. (1989) Synthesis and properties of 5-azido-UDP-glucose. *J. Biol. Chem.* 264, 11923-11933.
 - (71) Wallace, C. J. A. and Harris, D. E. (1984) The preparation of fully N- ϵ -acetimidylated cytochrome c. *Biochem. J.* 217, 589-594.
 - (72) Grossberg, A. L. and Pressman, D. (1963) Effect of acetylation on the active site of several antihapten antibodies: Further evidence for the presence of tyrosine in each site. *Biochemistry* 2, 90-96.
 - (73) Buttkus, H., Clark, J. R., and Feeney, R. E. (1965) Chemical modifications of amino groups of transferrins: Ovotransferrin, human serum transferrin and human lactotransferrin. *Biochemistry* 4, 998-1005.
 - (74) Haynes, R., Osuga, D. T., and Feeney, R. E. (1967) Modification of amino groups in inhibitors of proteolytic enzymes. *Biochemistry* 6, 541-547.
 - (75) Huynh, Q. K. (1988) Evidence for a reactive α -carboxyl group (Glu-418) at the herbicide glyphosate binding site of 5-enolpyruvylshikimate-3-phosphate synthase from *Escherichia coli*. *J. Biol. Chem.* 263, 11631-11635.
 - (76) Rogers, T. B., Borresen, T., and Feeney, R. E. (1978) Chemical modification of the arginines in transferrins. *Biochemistry* 17, 1105-1109.
 - (77) Riordan, J. E. (1973) Functional arginine residues in carboxypeptidase A—modification with butanedione. *Biochemistry* 12, 3915-3923.
 - (78) Yamasaki, R. B., Vega, A., and Feeney, R. E. (1980) Modification of available arginine residues in proteins by *p*-hydroxyphenylglyoxal. *Anal. Biochem.* 109, 32-40.
 - (79) Kasher, J. S., Allen, K. E., Kasamo, K., and Slayman, C. W. (1986) Characterization of an essential arginine residue in the plasma membrane H⁺-ATPase of *Neurospora crassa*. *J. Biol. Chem.* 261, 10808-10813.
 - (80) Melchior, W. B. and Fahrney, D. (1970) Ethoxyformylation of proteins. Reaction of ethoxyformic anhydride with α -chymotrypsin, pepsin and pancreatic ribonuclease at pH 4. *Biochemistry* 9, 251-258.
 - (81) Dominici, P., Tancini, B., and Voltattorni, C. B. (1985) Chemical modification of pig kidney 3,4-dihydroxyphenylalanine decarboxylase with diethyl pyrocarbonate. *J. Biol. Chem.* 260, 10583-10589.
 - (82) Spande, T. F. and Witkop, B. (1967) Tryptophan involvement in the function of enzymes and protein hormones as determined by selective oxidation with *N*-bromosuccinimide. *Methods Enzymol.* 11, 506-521.
 - (83) Horton, H. R. and Koshland, D. E. (1972) Modification of proteins with active benzylhalides. *Methods Enzymol.* 25, 468-482.
 - (84) Horton, H. R. and Koshland, D. E. (1967) Reactions with reactive alkylhalides. *Methods Enzymol.* 11, 556-565.
 - (85) Morrison, M. (1970) Iodination of tyrosine: Isolation of lactoperoxidase (bovine). *Methods Enzymol.* 17, 653-664.
 - (86) Sinn, H. J., Schrank, H. H., Friedrich, E. A., Via, D. P., and Dresel, H. A. (1988) Radioiodination of proteins and lipoproteins using *N*-bromosuccinimide as oxidizing agent. *Anal. Biochem.* 170, 186-192.
 - (87) Sokolovsky, M., Riordan, J. F., and Vallee, B. L. (1966) Tetranitromethane. A reagent for the nitration of tyrosyl residues in proteins. *Biochemistry* 5, 3582-3589.
 - (88) Brake, J. M. and Wold, F. (1962) Carboxymethylation of yeast enolase. *Biochemistry* 1, 386-391.
 - (89) Crestfield, A. M., Moore, S., and Stein, W. H. (1963) The preparation and enzymatic hydrolysis of reduced and S-carboxymethylated proteins. *J. Biol. Chem.* 238, 622-627.
 - (90) Markham, G. D. and Satishchandran, C. (1988) Identification of the reactive sulfhydryl groups of S-adenosylmethionine synthetase. *J. Biol. Chem.* 263, 8666-8670.
 - (91) Lewis, C. T., Seyer, J. M., and Carlson, G. M. (1989) Cysteine 288: An essential hyperreactive thiol of cytosolic phosphoenolpyruvate carboxykinase (GTP). *J. Biol. Chem.* 264, 27-33.
 - (92) Fujioka, M., Takata, Y., Konishi, K., and Ogawa, H. (1987) Function and reactivity of sulphhydryl groups of rat liver glycine methyltransferase. *Biochemistry* 26, 5696-5702.
 - (93) Stauffer, C. E. and Eison, D. (1969) The effect on subtilisin activity of oxidizing a methionine residue. *J. Biol. Chem.* 244, 5333-5338.
 - (94) Fujioka, M., Takata, Y., Konishi, K., and Ogawa, H. (1987) Function and reactivity of sulphhydryl groups of rat liver glycine methyltransferase. *Biochemistry* 26, 5696-5702.
 - (95) Degani, Y., Neumann, H., and Patchornik, A. (1970) Selective cyanylation of sulphhydryl groups. *J. Am. Chem. Soc.* 92, 6969-6976.
 - (96) Dagani, Y. and Patchornik, A. (1974) Cyanylation of sulphhydryl groups by 2-nitro-5-thiocyanobenzoic acid. High yield modification and cleavage of peptides at cysteine residues. *Biochemistry* 13, 1-11.
 - (97) Kindman, L. A. and Jencks, W. P. (1981) Modification and inactivation of CoA transferase by 5-nitro-5-(thiocyanato)benzoate. *Biochemistry* 20, 5183-5187.
 - (98) Smith, D. J. and Kenyon, G. L. (1974) Nonessentiality of the active sulphhydryl group of rabbit muscle creatine kinase. *J. Biol. Chem.* 249, 3317-3318.
 - (99) Hettinger, T. P. and Harbury, H. A. (1965) Guanidinated cytochrome c. *Biochemistry* 4, 2585-2589.
 - (100) Shields, G. S., Hill, R. L., and Smith, E. L. Preparation

- and properties of guanidinated mercuripapain. *J. Biol. Chem.* 234, 1747-1760.
- (101) Kaplan, H., Stevenson, K. J., and Hartley, B. S. (1971) Competitive labelling, a method for determining the reactivity of individual groups in proteins. *Biochem. J.* 124, 289-299.
- (102) Duggleby, K. G. and Kaplan, H. (1975) A competitive labeling method for the determination of the chemical properties of solitary functional groups in proteins. *Biochemistry* 14, 5168-5175.
- (103) Shewale, J. G. and Brew, K. (1982) Effects of Fe^{3+} binding on the microenvironments of individual amino groups in human serum transferrin as determined by differential kinetic labeling. *J. Biol. Chem.* 257, 9406-9415.
- (104) Rieder, R. and Bosshard, H. R. (1980) Comparison of the binding sites on cytochrome c for cytochrome c oxidase, cytochrome bc₁ and cytochrome c. Differential acetylation of lysyl residues in free and complexed cytochrome c. *J. Biol. Chem.* 255, 4732-4739.
- (105) Jackson, G. E. D. and Young, N. M. (1986) Determination of chemical properties of individual histidine and tyrosine residues of concanavalin A by competitive labelling with 1-fluoro-2,4-dinitrobenzene. *Biochemistry*, 25, 1657-1662.
- (106) Buechler, J. A., Vedvick, T. A., and Taylor, S. S. (1989) Differential labelling of the catalytic subunit of cAMP dependent protein kinase with acetic anhydride: substrate-induced conformational changes. *Biochemistry* 28, 3018-3024.
- (107) Ray, W. J. and Koshland, D. E. (1961) A method for characterizing the type and numbers of groups involved in enzyme action. *J. Biol. Chem.* 236, 1973-1979.
- (108) Redkar, V. D. and Kenkare, U. W. (1975) Effects of ligands on the reactivity of essential sulfhydryls in brain hexokinase. Possible interaction between substrate binding sites. *Biochemistry* 14, 4704-4712.
- (109) Horiike, K., Tsuge, H., and McCormick, D. B. (1979) Evidence for an essential histidyl residue at the active site of pyridoxamine(pyridoxine)-5'-phosphate oxidase from rabbit liver. *J. Biol. Chem.* 254, 6638-6643.
- (110) Jimenez, J. S., Kupfer, A., Gani, V., and Shaltiel, S. (1982) Conformational changes in the catalytic subunit of adenosine cyclic 3',5'-phosphate dependent protein kinase. Use for establishing a connection between one sulfhydryl group and the γ -subsite in the ATP site of this subunit. *Biochemistry* 21, 1623-1630.
- (111) Ogawa, H., Okamoto, M., and Fujioka, M. (1979) Chemical modification of the active site sulfhydryl group of saccharopine dehydrogenase (L-lysine-forming). *J. Biol. Chem.* 254, 7030-7035.
- (112) First, E. H. and Taylor, J. J. (1989) Selective modification of the catalytic subunit of cAMP-dependent protein kinase with sulfhydryl-specific fluorescent probes. *Biochemistry* 28, 3598-3605.
- (113) Makinen, A. L. and Nowak, T. (1989) A reactive cysteine in avian liver phosphoenolpyruvate carboxykinase. *J. Biol. Chem.* 264, 12148-12157.
- (114) Tsou, Chen-Lu (1962) Kinetic determination of essential side chains in proteins. *Sci. Sin.* 11, 1536-1558.
- (115) Kremer, A. B., Egan, R. M., and Sable, H. Z. (1980) The active site of transketolase two arginine residues are essential for activity. *J. Biol. Chem.* 255, 2405-2410.
- (116) Lakowicz, J. R. (1983) *Principles of Fluorescence Spectroscopy* Plenum Press, New York.
- (117) Stryer, L. and Haugland, R. P. (1967) Energy transfer: A spectroscopic ruler. *Proc. Natl. Acad. Sci.* 58, 719-726.
- (118) Stryer, L. (1978) Fluorescence energy transfer as a spectroscopic ruler. *Ann. Rev. Biochem.* 47, 819-846.
- (119) Haugland, R. P. (1989) *Molecular Probes Handbook of Fluorescent Probes and Research Chemicals* Molecular Probes, Inc., Eugene, OR.
- (120) Berliner, L. J. (1976) *Spin Labels* Academic Press, New York.
- (121) Wold, F. (1972) Bifunctional reagents. *Methods Enzymol.* 25, 623-651.
- (122) Wang, K. and Richards, F. M. (1974) An approach to nearest neighbor analysis of membrane proteins. *J. Biol. Chem.* 249, 8005-8018.
- (123) Uy, R. and Wold, F. (1977) Introduction of artificial crosslinks into proteins. *Adv. Exp. Med. Biol.* 86A, 169-186.
- (124) Das, M. and Fox, C. F. (1979) Chemical cross-linking in biology. *Annu. Rev. Biophys. Bioeng.* 8, 165-193.
- (125) Ji, T. H. (1983) Bifunctional reagents. *Methods Enzymol.* 91, 580-609.
- (126) Kennedy, J. F. and Cabral, J. M. S. (1983) In *Solid Phase Biochemistry* (W. H. Scouten, Ed.) pp 253-392, John Wiley, New York.
- (127) Laskin, A. I. (1985) *Enzymes and Immobilized Cells in Biotechnology* Benjamin/Cummings, Inc., Menlo Park, CA.
- (128) Hartmeir, W. (1986) *Immobilized Biocatalysts* Springer-Verlag, New York.
- (129) Mosbach, K. (1987) Immobilized enzymes and cells, part B. *Methods Enzymol.* 135.
- (130) Mosbach, K. (1987) Immobilized enzymes and cells, part C. *Methods Enzymol.* 136.
- (131) Weare, J. A. and Reichert, I. E. (1979) Studies with carbodiimide-cross-linked derivatives of bovine lutropin. I. The effects of specific group modifications on receptor site binding in testes. *J. Biol. Chem.* 254, 6964-6971.
- (132) Waldmeyer, B. and Bosshard, H. R. (1985) Structure of an electron transfer complex. I. Covalent cross-linking of cytochrome c peroxidase and cytochrome c. *J. Biol. Chem.* 260, 5184-5190.
- (133) Willing, A. H., Georgiadis, M. M., Rees, D. C., and Howard, J. B. (1989) Cross-linking of nitrogenase components structure and activity of the covalent complex. *J. Biol. Chem.* 264, 8499-8503.
- (134) Korodi, I., Asboth, B., and Polgar, L. (1986) Disulfide bond formation between the active-site thiol and one of the several free thiol groups of chymopapain. *Biochemistry* 25, 6895-6900.
- (135) Huston, E. E., Grammer, J. C., and Yount, R. G. (1988) Flexibility of the myosin heavy chain—Direct evidence that the region containing SH₁ and SH₂ can move 10 Å under the influence of nucleotide binding. *Biochemistry* 17, 8945-8952.
- (136) Hiratsuka, T. (1988) Cross-linking of three heavy chain domains of myosin adenosinetriphosphatase with a trifunctional alkylating reagent. *Biochemistry* 27, 4110-4114.
- (137) Peters, K. and Richards, F. M. (1977) Chemical cross-linking: Reagents and problems in studies of membrane structure. *Ann. Rev. Biochem.* 46, 523-551.
- (138) Davies, G. E. and Stark, G. R. (1970) Use of dimethylsuberimide, a cross-linking reagent, in studying the subunit structure of oligomeric proteins. *Proc. Natl. Acad. Sci. U.S.A.* 66, 651-656.
- (139) Dombradi, V., Hajdu, J., Bot, G., and Friedrich, P. (1980) Structural changes in glycogen phosphorylase as revealed by cross-linking with bifunctional diimides: phosphodephospho hybrid and phosphorylase a. *Biochemistry* 19, 2295-2299.
- (140) Pilch, P. E. and Czech, M. P. (1979) Interaction of cross-linking agents with the insulin effector system of isolated cells. *J. Biol. Chem.* 254, 3375-3381.
- (141) Staros, J. V., Lee, W. T., and Conrad, D. H. (1988) Membrane impermeant crosslinking reagents application to studies of the cell surface receptor for IgE. *Methods Enzymol.* 150, 503-512.
- (142) Heilman, H. D. and Holzner, M. (1981) The spatial organization of the active sites of the bifunctional oligomeric enzyme tryptophan synthetase: Crosslinking by a novel method. *Biochem. Biophys. Res. Commun.* 99, 1146-1152.
- (143) Sato, S. and Nakao, M. (1981) Cross-linking of intact erythrocyte membranes with a newly synthesized cleavable bifunctional reagent. *J. Biochem. (Tokyo)* 90, 1177-1181.
- (144) Moore, J. E. and Ward, W. H. (1956) Cross-linking of bovine plasma albumin and wool keratin. *J. Am. Chem. Soc.* 78, 2414-2418.

- (145) Hillel, Z. and Wu, C.-W. (1977) Subunit topography of RNA polymerase from *Escherichia coli*. A cross-linking study with bifunctional reagents. *Biochemistry* 16, 3334-3342.
- (146) Hingorani, V. N., Tobias, D. T., Henderson, J. T., and Ho, Y.-K. (1988) Chemical crosslinking of bovine retinal transducin and cGMP phosphodiesterase. *J. Biol. Chem.* 263, 6916-6926.
- (147) Kitagawa, T. and Aikawa, T. (1976) Enzyme coupled immunoassay of insulin using a novel coupling reagent. *J. Biochem. (Tokyo)* 79, 233-236.
- (148) Youle, R. J. and Neville, D. M. (1980) Anti-thy 1.2 monoclonal antibody linked to ricin is a potent cell-type-specific toxin. *Proc. Natl. Acad. Sci. U.S.A.* 77, 5483-5486.
- (149) Yoshitake, S., Yamada, Y., Ishikawa, E., and Masseyeff, R. (1979) Conjugation of glucose oxidase from *Aspergillus niger* and rabbit antibodies using *N*-hydroxysuccinimide ester of *N*-(4-carboxycyclohexylmethyl)maleimide. *Eur. J. Biochem.* 101, 395-399.
- (150) Lambert, J. M., Senter, P. D., Young, A. Y. Y., Blattler, W. A., and Goldmacher, V. S. (1985) Purified immunotoxins that are reactive with human lymphoid cells. *J. Biol. Chem.* 260, 12035-12041.
- (151) Carlsson, J., Dreyen, H., and Axen, R. (1978) Protein thiolation and reversible protein-protein conjugation *N*-succinimidyl 3(2-pyridylthio) propionate, a new heterobifunctional reagent. *Biochem. J.* 173, 723-737.
- (152) O'Keefe, D. O. and Draper, R. K. (1985) Characterization of a transferrin-diphtheria conjugate. *J. Biol. Chem.* 260, 932-937.
- (153) Jue, R., Lambert, J. M., Pierce, L. R., and Traut, R. R. (1978) Addition of sulfhydryl groups to *Escherichia coli* ribosomes by protein modification with 2-iminothiolane (methyl 4-mercaptobutyrimide). *Biochemistry* 17, 5399-5406.
- (154) Marsh, J. W. (1988) Antibody-mediated routing of diphtheria toxin in murine cells results in a highly efficacious immunotoxin. *J. Biol. Chem.* 263, 15993-15999.
- (155) Cover, J. R., Lambert, J. M., Norman, C. M., and Traut, R. R. (1981) Identification of proteins at the subunit interface of the *Escherichia coli* ribosome by cross-linking with dimethyl 3,3'-dithiobis(propionimidate). *Biochemistry* 20, 2843-2852.
- (156) Staros, J. V., Wright, R. W., and Swingle, D. M. (1986) Enhancement by *N*-hydroxysulfosuccinimide of water-soluble carbodiimide modified coupling reactions. *Anal. Biochem.* 156, 220-222.
- (157) Koyama, Y. and Taniguchi, A. (1986) Studies on chitin X. Homogeneous cross-linking of chitosan for enhanced cupric ion adsorption. *J. Appl. Polymer Sci.* 31, 1951-1954.
- (158) Golander, C.-G. and Eriksson, J. C. (1987) ESCA studies of the adsorption of polyethyleneimine and glutaraldehyde-reacted polyethyleneimine on polyethylene and mica surfaces. *J. Colloid Interface Sci.* 119, 38-48.
- (159) Korn, A. H., Fearheller, S. H., and Filachione, E. M., Glutaraldehyde: Nature of the reagent. *J. Mol. Biol.* 65, 525-529.
- (160) Kirkeby, S., Jakobsen, P., and Moe, D. (1987) Glutaraldehyde—"pure and impure". A spectroscopic investigation of two commercial glutaraldehyde solutions and their reaction products with amino acids. *Anal. Lett.* 20, 303-315.
- (161) Gregory, J. D. (1955) The stability of *N*-ethylmaleimide and its reaction with sulfhydryl groups. *J. Am. Chem. Soc.* 77, 3922-3923.
- (162) Knight, P. (1979) Hydrolysis of *p*-*N,N'*-phenylenebismaleimide and its adducts with cysteine. *Biochem. J.* 179, 191-197.
- (163) Staros, J. V. (1982) *N*-Hydroxysulfosuccinimide active esters: Bis(*N*-hydroxysulfosuccinimide) esters of two dicarboxylic acids are hydrophilic, membrane-impermeant, protein cross-linkers. *Biochemistry* 21, 3940-3955.
- (164) Yoshitake, S., Imagawa, M., Ishikawa, E., Niitsu, Y., Urushizaki, I., Nishiura, M., Kanazawa, R., Kurosaki, H., Tachibana, S., Nakazawa, N., and Ogawa, H. (1982) Mild and efficient conjugation of rabbit Fab' and horseradish peroxidase using a maleimide compound and its use for enzyme immunoassay. *J. Biochem.* 92, 1413-1424.
- (165) Galaray, R. E., Craig, L. C., Jamieson, J. D., and Printz, M. P. (1974) Photoaffinity labeling of peptide hormone binding sites. *J. Biol. Chem.* 249, 3510-3518.
- (166) Wood, C. L. and O'Dorisio, M. S. (1985) Covalent cross-linking of vasoactive intestinal polypeptide to its receptors on intact human lymphoblasts. *J. Biol. Chem.* 260, 1243-1247.
- (167) Schmitt, M., Painter, R. G., Jesaitis, A. J., Preissner, K., Sklar, L. A., and Cochran, C. G. (1983) Photoaffinity labeling of the *N*-formyl peptide receptor binding site of intact human polymorphonuclear leukocytes. *J. Biol. Chem.* 258, 649-654.
- (168) Benesch, R. and Benesch, R. E. (1956) Formation of peptide bonds by aminolysis of homocysteine thiolactones. *J. Am. Chem. Soc.* 78, 1597-1599.
- (169) Klotz, I. M. and Heiney, K. E. (1962) Introduction of sulfhydryl groups into proteins using acetylmercaptosuccinic anhydride. *Arch. Biochem. Biophys.* 96, 605-612.
- (170) Julian, R., Duncan, S., Weston, P. D., and Wrigglesworth, R. (1983) A new reagent which may be used to introduce sulfhydryl groups into proteins, and its use in the preparation of conjugates for immunoassay. *Anal. Biochem.* 132, 68-73.
- (171) Gitman, A. G., Kahane, I., and Loyter, A. (1985) Use of virus-attached antibodies or insulin molecules to mediate fusion between sendai virus envelopes and neuraminidase-treated cells. *Biochemistry* 24, 2762-2768.
- (172) Gordon, R. D., Fieles, W. E., Schotland, D. L., Hogue-Angeletti, R., and Barchi, R. L. (1987) Topographical localization of the C-terminal regions of the voltage-dependent sodium channel from *Electrophorus electricus* using antibodies raised against a synthetic peptide. *Proc. Natl. Acad. Sci. U.S.A.* 84, 308-312.
- (173) Senter, P. D., Saulnier, M. G., Schreiber, G. J., Hirschberg, D. L., Brown, J. P., Hellstrom, I., and Hellstrom, K. E. (1988) Anti-tumor effects of antibody-alkaline phosphatase conjugates in combination with etoposide phosphate. *Proc. Natl. Acad. Sci. U.S.A.* 85, 4842-4846.
- (174) Hirs, C. H. W. (1967) Protein structure. *Methods Enzymol.* 11.
- (175) Hirs, C. H. W. and Timasheff, S. N. (1983) Enzyme structure, part I. *Methods Enzymol.* 91.
- (176) Baker, B. R. (1967) *Design of Active-Site-Directed Irreversible Enzyme Inhibitors* Wiley-Interscience, New York.
- (177) Means, G. E. and Feeney, R. E. (1971) *Chemical Modification of Proteins* Holden-Day, San Francisco, CA.
- (178) Glazer, A. N., Delange, R. J., and Sigman, D. S. (1975) *Chemical Modification of Proteins. Laboratory Techniques in Biochemistry and Molecular Biology* (T. S. Work and E. Work, Eds.) American Elsevier Publishing Co., New York.
- (179) Lundblad, R. L. and Noyes, C. M. (1984) *Chemical Reagents for Protein Modification* Vols. 1 and 2, CRC Press, Boca Raton, FL.
- (180) Widder, K. J. and Green, R. (1985) Drug and enzyme targeting, Part A. *Methods Enzymol.* 112.
- (181) Pfeleiderer, G. (1985) Chemical Modifications of Proteins. In *Modern Methods in Protein Chemistry* (H. Tschesche, Ed.) Walter de Gruyter, Berlin and New York.
- (182) Feeney, R. E. (1987) Chemical modification of proteins: Comments and perspectives. *Int. J. Pept. Protein Res.* 27, 145-161.
- (183) Eyzaguirro, J. (1987) *Chemical Modification of Enzymes: Active Site Studies* John Wiley and Sons, New York.

Exhibit 24

Re-engineering of Human Urokinase Provides a System for Structure-based Drug Design at High Resolution and Reveals a Novel Structural Subsite*

(Received for publication, September 17, 1999, and in revised form, November 30, 1999)

Vicki Nienaber[‡], Jieyi Wang[¶], Don Davidson[¶], and Jack Henkin[¶]

From the Departments of [‡]Structural Biology and [¶]Cancer Research, Abbott Laboratories, Abbott Park, Illinois 60064

Inhibition of urokinase has been shown to slow tumor growth and metastasis. To utilize structure-based drug design, human urokinase was re-engineered to provide a more optimal crystal form. The redesigned protein consists of residues Ile¹⁶-Lys²⁴³ (in the chymotrypsin numbering system; for the urokinase numbering system it is Ile¹⁵⁹-Lys⁴⁰⁴) and two point mutations, C122A and N145Q (C279A and N302Q). The protein yields crystals that diffract to ultra-high resolution at a synchrotron source. The native structure has been refined to 1.5 Å resolution. This new crystal form contains an accessible active site that facilitates compound soaking, which was used to determine the co-crystal structures of urokinase in complex with the small molecule inhibitors amiloride, 4-iodo-benzo(b)thiophene-2-carboxamide and phenyl-guanidine at 2.0–2.2 Å resolution. All three inhibitors bind at the primary binding pocket of urokinase. The structures of amiloride and 4-iodo-benzo(b)thiophene-2-carboxamide also reveal that each of their halogen atoms are bound at a novel structural subsite adjacent to the primary binding pocket. This site consists of residues Gly²¹⁸, Ser¹⁴⁶, and Cys¹⁹¹-Cys²²⁰ and the side chain of Lys¹⁴³. This pocket could be utilized in future drug design efforts. Crystal structures of these three inhibitors in complex with urokinase reveal strategies for the design of more potent nonpeptidic urokinase inhibitors.

Cancer cell invasion, the spread and growth of tumor metastases, is a primary cause of mortality and morbidity of malignancy (2), and this invasion requires the degradation of basement membranes and other extracellular protein structures. Urokinase has been shown to be strongly associated with tumor cells (3) and to play a role in basement membrane degradation via a cascade mechanism involving activation of plasminogen and the metalloproteases (4–6). Furthermore, inhibitors of urokinase have been reported to slow tumor metastasis as well as growth of the primary tumor (7–15). These inhibitors include the small molecules 4-iodo-benzo(b)thiophene-2-carboxamide (B428),¹ 4-benzodioxolanylethyl benzo(b)thiophene-2-carboxamide (B623) (12–14), and amiloride (8, 15). These compounds are competitive inhibitors of uroki-

nase and have been proposed to bind at the primary binding pocket common to all trypsin-like serine proteases (15). However, none of these compounds possess all of the characteristics of a good therapeutic agent for the treatment of cancer.

Structure-based drug design has become an important tool for improving the potency and pharmacological characteristics of compounds toward providing therapeutic agents. This method has contributed to the development of potent and specific inhibitors for many targets such as HIV protease, cyclooxygenase-2, influenza neuraminidase, and the metalloproteases (16–22). To most efficiently apply crystallography-driven structure-based drug design, it is preferable that the crystals have certain properties. One property is that active site of the target is open in the crystal lattice. This molecular packing permits the diffusion and binding of compounds into the active site and eliminates the need to optimize crystal growth in the presence of each inhibitor. Another important property is that the crystals reproducibly diffract to high resolution (2.5–2.0 Å). It is preferable that this data quality is achievable on a conventional rotating anode source, thereby eliminating the need for travel to synchrotron facilities. The higher resolution data facilitate unambiguous map interpretation and minimize the average atomic positional error (23). Hence, an appropriate crystal form can greatly facilitate the process of structure-based drug design. A crystal system exists for urokinase, although it does not fully encompass the preferred properties outlined above.

Human low molecular weight (LMW) urokinase has been crystallized in complex with the peptidic inhibitor Glu-Gly-Arg-chloromethyl ketone (1). This structure reveals the geometry of the urokinase active site as well as the orientation of a peptide inhibitor in the substrate-binding groove. However, the LMW urokinase crystals diffract to lower resolution (2.5 Å resolution, synchrotron radiation; 3.0 Å resolution, rotating anode source) and utilize co-crystallization to achieve the target-ligand complex. In addition, the active site is in close contact with another molecule because of a noncrystallographic 2-fold axis near the active site. This interaction could limit minor ligand induced conformational shifts and perhaps distort the active site conformation. Furthermore, the noncrystallographic and crystallographic packing effectively blocks the active site such that it would be difficult to diffuse small molecules into the active site in this crystal form (if they were not blocked by the irreversible covalent inhibitor). Hence, although this system may be used for modeling of small molecule urokinase inhibitors, it may not provide an ideal system for structure-based drug design. Therefore, to design an anti-cancer therapeutic, a new crystal form of human urokinase was sought to facilitate the application of structure-based drug design. The strategy utilized protein engineering and information from the reported LMW urokinase structure to design an altered protein sequence to yield a new crystal form.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: Dept. of Structural Biology, Abbott Laboratories, D46Y/AP10-LL, 100 Abbott Park Rd., Abbott Park, IL 60064-6098. Tel.: 847-935-0918; Fax: 847-937-2625; E-mail: vicki.nienaber@abbott.com.

¹ The abbreviations used are: B428, 4-iodo-benzo(b)thiophene-2-carboxamide; B623, 4-benzodioxolanylethyl benzo(b)thiophene-2-carboxamide; LMW, low molecular weight; S2444, H-D-pyroglyutamyl-Gly-L-Arg-p-nitroanilide.

The new form of urokinase, micro-urokinase, crystallizes under conditions very similar to the low molecular weight form (1), although crystal packing and data quality are very different. This new crystal form contains a monomer in the asymmetric unit and diffracts to ultra-high resolution ($d_{\min} = 1.03$ Å). In addition, this crystal form has an open active site permitting direct diffusion of compounds into the apo-crystals and is therefore ideal for providing precise structure determinations for urokinase ligand complexes by the soaking technique.

The re-engineered crystal system and soaking technique were utilized to determine the co-crystal structure of urokinase in complex with a series of small molecule inhibitors at 2.0 or 2.2 Å resolution. Two of these inhibitors, amiloride (24), and B428 (25, 26), have been shown to reduce tumor size and metastasis (8, 12–15), whereas the effect of the third, phenylguanidine (27) has not been reported to date. These complex structures were completed to determine the binding orientation of each compound to urokinase. This information in turn may be utilized to design molecules of increased potency toward discovery of an anti-cancer therapeutic compound.

EXPERIMENTAL PROCEDURES

Recombinant Micro-urokinase—Micro-urokinase was engineered by polymerase chain reaction manipulations using a human urokinase cDNA as a template (28). The C279A and N302Q mutations were made by the method of polymerase chain reaction based site-directed mutagenesis. Urokinase native leader sequence was fused directly to Ile¹⁶⁹ by polymerase chain reaction. This product was ligated to a baculovirus transfer vector pJVP10z (29). The final expression vector sequence was confirmed by DNA sequencing.

The pJVP10z-micro-urokinase vector was transfected into Sf9 cells by the calcium phosphate precipitation method using the BaculoGold kit from Pharmingen (San Diego, CA). Single recombinant virus expressing micro-urokinase was plaque purified by standard methods, and a large stock of the virus was prepared. Large scale expression of micro-urokinase was performed in suspension in High-Five cells, (Invitrogen, San Diego, CA) growing in Excell 405 serum free medium (JRH Biosciences, Lenexa, KS) at 27 °C. Urokinase activity in the supernatant was measured by amidolysis of the chromogenic urokinase substrate H-D-pyrroglutamyl-Gly-L-Arg-p-nitroanilide (S2444; Helena Laboratories, Beaumont, TX). The culture supernatant was harvested as the starting material for purification. Protease inhibitors, iodoacetamide (10 mM), benzamide (5 mM), and EDTA (1 mM) were added to the pooled culture medium. The medium was diluted 5-fold with 5 mM HEPES, pH 7.5, and filtered through 1.2 and 0.2-μm membranes. The micro-urokinase protein was captured onto Sartorius membrane adsorbent S100 (Sartorius, Edgewood, NY) by passing the medium through the membrane at a flow rate of 50–100 ml/min. After extensive washing with 10 mM HEPES, pH 7.5, containing 10 mM iodoacetamide, 5 mM benzamide, and 1 mM EDTA, micro-urokinase was eluted from S100 membrane with a NaCl gradient (20–500 mM, 200 ml) in 10 mM HEPES buffer, pH 7.5, 10 mM iodoacetamide, 5 mM benzamide, 1 mM EDTA. The eluate was diluted 10-fold with the above 10 mM HEPES buffer containing inhibitors, and loaded onto a S20 column (Bio-Rad). Micro-urokinase was eluted with a 20× column volume NaCl gradient (20–500 mM). No inhibitors were used in the elution buffers. The eluate was then diluted 5-fold with 10 mM HEPES buffer, pH 7.5, and loaded onto a heparin-agarose (Sigma) column. Micro-urokinase was eluted with a NaCl gradient from 10–250 mM. The heparin column eluate of micro-urokinase was applied to a benzamide-agarose (Sigma) column equilibrated with 10 mM HEPES buffer, pH 7.5, 200 mM NaCl. The column was washed with the equilibration buffer, and the urokinase was eluted with 50 mM NaOAc, pH 4.5, 500 mM NaCl. The micro-urokinase eluate was concentrated to 4 ml by ultrafiltration and applied to a Sephadex G-75 column equilibrated with 20 mM NaOAc, pH 4.5, 100 mM NaCl. The single peak containing micro-urokinase was collected and lyophilized as the final product.

Amidolytic Kinetics of Urokinase and Micro-urokinase—The effects of synthetic inhibitors on the steady state amidolytic activity of LMW urokinase or micro-urokinase toward the chromogenic substrate, S2444 (Helena Laboratories), was characterized by the formation of p-nitroaniline (30). Briefly, 0–50 μM concentration of inhibitors were tested against 25 IU/ml (0.14 ng/ml) LMW urokinase or micro-urokinase and 0.4–4.0 mM concentrations of S2444 in 200 μl volumes in phosphate-

buffered saline and 0.01% bovine serum albumin, pH 7.4. Incubations were performed at 37 °C with absorbance at 405 nm recorded every 11 s for 20 min. Data were plotted as $1/S$ versus $1/v$ for Lineweaver-Burk analysis and the calculation of inhibition constants. K_i values were obtained from replots of the resultant slopes versus I (26, 31).

Protein Crystallography—Crystals were obtained by the hanging drop vapor diffusion method. A typical well solution of 0.15 M Li_2SO_4 , 20% polyethylene glycol MW 4000 in succinate buffer, pH 4.8–6.0, was used. On the coverslip, 2 μl of well solution is mixed with 2 μl of protein solution, and the slip is sealed over the well. Crystallization occurred at 18–24 °C within 24 h. The protein solution was composed of 6 mg/ml (0.21 mM) micro-urokinase in 10 mM citrate, pH 4.0, 3 mM ϵ -amino caproic acid *p*-carboxyphenyl ester chloride with 1% Me_2SO co-solvent. The resultant micro-urokinase crystals are composed of enzyme with an empty active site. The compound ϵ -amino caproic acid *p*-carboxyphenyl ester chloride is reported to inhibit urokinase with an apparent K_i of 0.3 μM at neutral pH and was co-crystallized with urokinase in an attempt to obtain a complex structure (32). Repeated tests with this compound resulted in a structure with an active site occupied only by ordered solvent molecules even at 1.5 Å resolution. Hence, we have hypothesized that this inhibitor is degraded during the crystallization experiment albeit critical for obtaining urokinase crystals. Studies are underway to try to understand the mechanism of this phenomenon.

The micro-urokinase crystals belong to the space group $P2_12_12_1$, with unit cell dimensions of $a = 55.16$ Å, $b = 53.00$ Å, $c = 82.30$ Å and $\alpha = \beta = \gamma = 90^\circ$ and diffract beyond 1.5 Å on a Rigaku RTP 300 RC rotating anode source equipped with an RAXISII detector. In addition, a 1.03 Å resolution native data set was collected on a CCD detector at beam line F1 of the Cornell High Energy Synchrotron Source in Ithaca, NY. All data were collected at 100–160 K and processed by the program package DENZO (33). Before crystals were frozen, they were passed through a solution of 0.15 M Li_2SO_4 , 20% polyethylene glycol MW 4000, succinate buffer, pH 4.8–6.0, and 20% glycerol for cryogenic protection. Data were collected at low temperature to preserve the diffraction of the crystal throughout data acquisition. The crystal structure was determined by the molecular replacement method using the program AMORE (34). The LMW urokinase structure was used as the search probe (1) (Protein Data Bank entry 1LMW) against the RAXISII data.

The structure was refined to 1.5 Å resolution using the synchrotron data and the program package XPLOR (35) by a combination of rigid body, simulated annealing maximum likelihood refinement, and maximum likelihood positional refinement. Electron density maps to 1.5 Å resolution were inspected on a Silicon Graphics INDIGO2 workstation using the program package QUANTA 97 (Molecular Simulations, Inc.). At 1.5 Å resolution constrained individual temperature factor refinement was also included in the refinement cycle. Electron density maps to 1.5 Å resolution were examined, and water molecules and bound ions were identified as positive peaks in the $F_o - F_c$ map at least 4 σ above noise. Refinement continued with automatic water addition using the XWAT feature of SHELXL (36). Final refinement steps included cycles of model building where disorder and additional solvent molecules were added. The final R-factor is 19.2% with a R_{free} of 21.8%.

To obtain the amiloride, B428, or phenylguanidine micro-urokinase complex structures, crystals of urokinase were placed in 50 μl of crystallization mother liquor to which 0.5 μl of a 1 mg/10 μl compound solution was added. The solid compound was obtained from the Abbott chemical repository and was initially dissolved in Me_2SO . Crystals were allowed to incubate for 12–15 h at 24 °C and prepared for data collection in a manner identical to that of the native crystals. Data were collected on a Rigaku RTP 300 RC rotating anode source equipped with an RAXISII detector at 160 K by the method of flash freezing. Data were processed using the HKL program suite (33). Initial electron density maps were calculated using the program package XPLOR (35) and the 1.5 Å native model. All electron density maps were inspected on a Silicon Graphics INDIGO2 workstation using QUANTA 97, and the orientation of all compounds were clearly visualized in the initial $2F_o - F_c$ map. The complexes were refined to 2.0 Å resolution using the program package XPLOR. Refinement consisted of alternating steps of positional and B-factor refinement. Ordered solvent molecules were identified as positive peaks in the $F_o - F_c$ map that were 4 σ above noise.

Table I summarizes statistics for all micro-urokinase models. All data are between 89 and 90% complete with a merging R_{sym} between 7 and 11% and an I/σ between 12 and 15. The native model is refined to a R_{factor} of 19.2% and R_{free} of 21.8% at 1.5 Å resolution. The overall B-factor for the protein is 12 Å², and the overall B-factor for the 337 ordered solvent molecules is 26 Å². The current native model also

TABLE I
Data quality statistics

	Complete %	I/I _r	R _{sym} (square) ^a	R _{factor} ^b	R _{free} ^c
Native					
Overall	96.6	15	0.075	19.1	21.8
1.53–1.50 Å	95.3	9	0.113	21.2 (1.57–1.50)	25.8 (1.57–1.50)
B428					
Overall	89.9	16.8	0.083	20.9	27.7
2.05–2.0 Å	88.4	5	0.203	20.0	29.4
Amiloride					
Overall	99.8	12.4	0.108	21.5	29.1
2.3–2.2	99.8	4.3	0.358	19.1	26.9
Phenyl guanidine					
Overall	90.3	13.5	0.086	18.9	22.1
2.06–2.00	94.2	4.5	0.254	24.3	24.8

^a $R_{\text{sym}} = \sum (|I - \langle I \rangle|) / \sum I$ ^b $R_{\text{factor}} = \sum |F_o - F_c| / \sum F_o$ ^c Value of the R_{factor} where 10% of the data were randomly removed from the refinement.

contains three ordered sulfate ions, and two alternate side chain conformations located at the active site. All backbone atoms are well defined in the final $2F_o - F_c$ map with atomic B-factors at or below 30 Å². The B428 model is refined to 2.0 Å resolution with a R_{factor} of 20.9% and a R_{free} of 27.7%, while the amiloride model is refined to 2.2 Å resolution with a R_{factor} of 21.5% and a R_{free} of 29.1%. The phenylguanidine model is refined to 2.0 Å resolution with a R_{factor} of 18.9% and a R_{free} of 22.1%. Data for the complex structures were of quality comparable with that of native structures collected under the same conditions on a rotating anode source.

RESULTS

Redesign of LMW Urokinase—To redesign the LMW urokinase sequence for the purpose of improving the crystal characteristics, the LMW urokinase coordinate file (Protein Data Bank entry 1LMW) was examined for sequences of excessively high B-factor, suggesting areas of disorder. The hypothesis is that areas of high disorder in the structure may contribute to the overall disorder of the crystals and/or may interfere with optimal crystal packing. The LMW urokinase structure consists of residues 136–158 of the A-chain and 159–411 of the B-chain connected by a disulfide bridge between Cys¹⁴⁸ and Cys²⁷⁹ (urokinase numbering).² The B-chain corresponds to the serine protease domain, whereas the 21 residue A-chain lacks the kringle and epidermal growth factor domains present in full-length urokinase. The A-chain is reported to be an area of high disorder (1), and examination of the protein data bank coordinate file (Protein Data Bank entry 1LMW) reveals that residues 148–155 of the A-chain have an average B-factor of 64 Å² ranging from 26 Å² for the disulfide-linked sulfur of residue Cys¹⁴⁸ to 110 Å² for Pro¹⁵⁵. The very high B-factors for the LMW urokinase A-chain confirm this observation. Consequently, the A-chain was removed as a first step in the redesign. Furthermore, to remove the resultant free thiol on the B-chain, Cys¹⁴⁸ was mutated to an alanine.

Further examination of the LMW urokinase coordinate file indicates a second area of disorder consisting of residues 405–411 of the C terminus where the average B-factor is 147 Å². Residues 407–411 represent a five residue extension in urokinase relative to other trypsin-like serine proteases. However, because residues 405–406 also have high atomic B-factors, the entire 405–411 segment was removed. The final potential site for disorder is the glycosylation site at residue 302. This glycosylation site was removed by an N302Q mutation to facilitate expression of the glycosylation-free protein in baculovirus. Hence, the re-engineered urokinase (micro-urokinase) consists

of residues Ile¹⁵⁹–Lys⁴⁰⁴ (Ile¹⁶–Lys²⁴³ chymotrypsin numbering system) with the two point mutations C279A (C122A) and N302Q (N145Q).

Micro-urokinase Crystal Packing—Micro-urokinase crystallizes with a monomer in the asymmetric unit (P2₁,2₁), whereas the LMW urokinase crystal form has a dimer in the asymmetric unit (R3) with intimate contacts at the substrate-binding site. Specifically, in LMW urokinase, residues 94–101 from each molecule (chymotrypsin numbering system as aligned by Ref. 1)² form a series of intermolecular main chain hydrogen bonds resulting in an extended four stranded β-sheet (1). From the LMW urokinase structure, it was seen that this loop decreases the size of the S₄ pocket relative to that at the substrate-binding site of other serine proteases such as thrombin, Factor Xa and tissue plasminogen activator (1, 37–39). Hence, this loop provides a critical structural feature of the substrate-binding groove. However, because of the close crystal contact at this site in the LMW urokinase crystals, the possibility existed that the structure of the substrate-binding site may be distorted or conformationally restricted. The new crystal form of micro-urokinase lacks the close crystal contact present in LMW urokinase, and an overlay of the two structures indicates that the conformation of this loop is essentially identical in the two crystal forms. Consequently, it is unlikely that packing in either crystal system affects the conformation of this loop and the resultant shape of the S₄ pocket, although the more open micro-urokinase packing may allow for inhibitor-induced conformational shifts.

Examination of crystal packing at the A-chain-binding cleft gives insight into why micro-urokinase yields different lattice packing and better diffracting crystals (a sample of the final $2F_o - F_c$ electron density map at 1.5 Å resolution is shown in Fig. 1A). In LMW urokinase, the A-chain binds in a cleft composed of residues 25–29, 116–122, and 201–208. In the crystal structure of micro-urokinase, there is no A-chain, and the A-chain-binding cleft is partially occupied by a symmetry related molecule. Specifically, a hydrophobic loop extending from 144 to 150 in the symmetry related molecule is directly bound at the A-chain site such that Tyr¹⁴⁹-OH of the loop is involved in two hydrogen bonds at the A-chain cleft (Ser²⁰²-N and Ser¹³⁵-O). In LMW urokinase, the A-chain blocks this set of interactions. Thus, in micro-urokinase, removal of the A-chain exposes a new “binding site” for the 144–150 loop of another micro-urokinase molecule permitting a new lattice to form. This interaction at the A-chain cleft probably contributes to the improved crystal quality by being both a site of nucleation as well as by facilitating very close contact between adjacent molecules.

² The urokinase numbering system is used for discussion of the sequence re-engineering work, whereas the chymotrypsin numbering system as aligned by Ref. 1 is used for discussion of the serine protease domain structure for micro-urokinase.

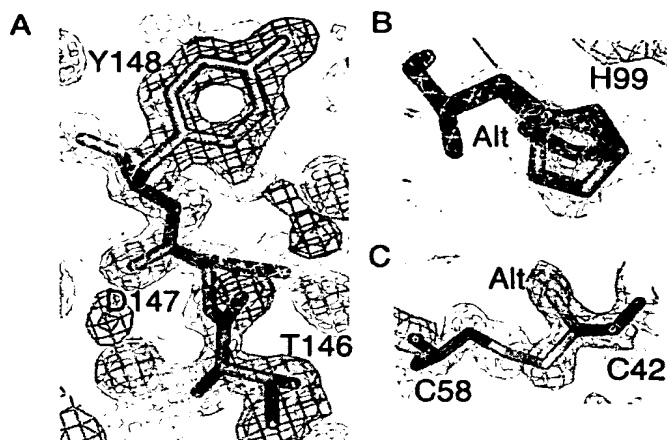


FIG. 1. A, final $2F_o - F_c$ electron density map contoured at 1σ for native micro-urokinase at 1.5 Å resolution. Residues 146–148 are depicted in thick lines. B, $2F_o - F_c$ (purple) and $F_o - F_c$ (green) at His⁹⁹. The $2F_o - F_c$ map is contoured at 1σ , and the $F_o - F_c$ is contoured at 3σ . The map is for refinement of the side chain in one conformation. C, $2F_o - F_c$ (purple) and $F_o - F_c$ (green) at Cys⁴². The $2F_o - F_c$ map is contoured at 1σ , and the $F_o - F_c$ is contoured at 3σ . The map is for refinement of the side chain in one conformation.

Micro-urokinase and LMW urokinase are nearly identical in structure (overall rms deviation for main chain atoms, 0.8 Å) with one significant structural change near a site of re-engineering. As discussed above, removal of the A-chain results in an empty cavity. One loop (201–210) forming this site undergoes a conformational shift relative to LMW urokinase with rms deviation (main chain) ranging from 1.1 to 1.8 Å with the largest shift being for Arg²⁰⁶. However, although this loop is involved in a crystal packing interaction, the conformation of the 144–150 of the symmetry related molecule is the same for both micro-urokinase and LMW urokinase. Other sites of variation include the flexible loop at residues 37–37D (rms deviation main chain, 1.7 – 3.5 Å), residues 17–19 (rms deviation main chain, 1.1 – 2.1 Å) and residues 185B–186 (rms deviation main chain, 1.7 Å). All areas were of high b-factor in the LMW urokinase structure (b-factor > 60 – 90 Å^2) but of significantly lower b-factor in the micro-urokinase structure (b-factor $< 20\text{ Å}^2$) with the exception of residues 17–19, which were of low b-factors in both structures. The 17–19 segment was clearly defined in the final $2F_o - F_c$ electron density maps of micro-urokinase and is not near any re-engineered sites. Residues 185B–186 were remodeled in the higher-resolution structure. In the lower resolution LMW urokinase structure, Trp¹⁸⁶ was exposed to solvent and Gln^{185B} was buried. The higher resolution data clearly placed Trp¹⁸⁶ in the protein core with Gln^{185B} exposed to solvent.

Active Site of Native Micro-urokinase—Like the overall molecular fold, the active sites of LMW urokinase and micro-urokinase are nearly identical (rms deviation, $< 0.8\text{ Å}$). The higher resolution data did not depict any large side chain movements relative to LMW urokinase but did show an alternate side chain conformation for two residues (Fig. 1, B and C) in addition to a bound sulfate ion (see Fig. 3C). The sulfate ion is bound near the oxyanion hole (40), where O1 is accepting hydrogen bonds from Gly¹⁹³-NH (2.8 Å) and Ser¹⁹⁵-OH (2.8 Å), whereas O₂ is accepting a hydrogen bond from His⁵⁷-Ne2 (2.8 Å). Hence, the higher resolution data revealed more structural details at the active site.

In Fig. 1B, native 1.5 Å $2F_o - F_c$ (contoured at 1σ) and $F_o - F_c$ (contoured at 3σ) electron density maps depict that the side chain of His⁹⁹ is in multiple conformations. These maps were calculated before the alternate conformation had been included

TABLE II
Inhibition constants determined for LMW urokinase and micro-urokinase
Ring numbering is shown in conjunction with the chemical structure for each inhibitor.

	LMW-urokinase	Ki (μM)	micro-urokinase
<p>B428</p>	0.490 ± 0.018	0.512 ± 0.022	
<p>Amiloride</p>	7.2 ± 0.2	6.9 ± 0.4	
<p>Phenylguanidine</p>	20.6 ± 1.0	17.4 ± 1.1	

in the model. As presented in Fig. 1B, one His⁹⁹ conformation is identical to that observed with LMW urokinase. In this conformation, His⁹⁹-Nδ1 accepts a hydrogen bond from Tyr⁹⁴-OH (2.9 Å). In the alternate conformation (modeled into the green positive peak; Fig. 1B), the His⁹⁹ imidazole is rotated approximately 90° about the Cβ-Cγ bond resulting in a different hydrogen bonding pattern. Here, His⁹⁹-Nδ1 can donate a hydrogen bond to Asp¹⁰²-Oδ1 (3.2 Å). The His⁹⁹ side chain forms part of both the S₄ and S₂ pockets. Hence, a change in the conformation of His⁹⁹ results in a change in the overall shape of S₂ and S₄, suggesting that the side chain movement would effect a drug design strategy directed toward the substrate-binding groove.

The side chain of Cys⁴² is also observed in two side chain conformations and is near the active site (Fig. 1C). In what is likely the major conformation, the Cys⁴²-Cys⁵⁸ disulfide bridge is intact. However, in the alternate conformation, the disulfide is broken and the Cys⁴² thiol group lies in a small hydrophobic pocket formed by the side chains of Phe⁵⁹, Ile²⁹, and Val⁴¹. This side chain shift is unexpected as the Cys⁴²-Cys⁵⁸ disulfide bridge is present all trypsin-like serine protease structures, and its proximity to the catalytic triad suggests that it may structurally stabilize the active site. Hence, one might expect the catalytic activity to be affected when this disulfide bridge is broken. On the other hand, one must note that this observation occurs in the solid state and that further solution work would be necessary to determine its physiological significance.

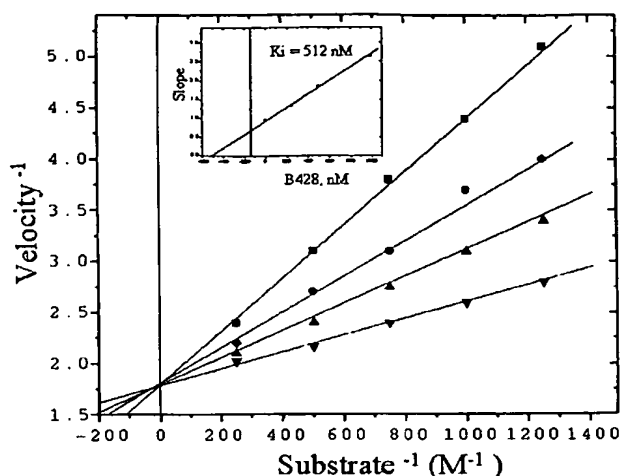


FIG. 2. Lineweaver-Burke analyses of B428 inhibition of micro-urokinase were performed in amidolytic chromogenic assays with S2444 as described under "Experimental Procedures." S2444 substrate concentrations were 0.8, 1.0, 1.3, 2.0, and 4.0 mM. B428 concentrations were 0 nM (∇), 250 nM (Δ), 500 nM (\bullet), and 1000 nM (\blacksquare). Data represent the means of triplicate determinations. K_i values were determined by replots of slope versus inhibitor concentration (inset) and are represented in Table II.

Examination of crystal packing at the active site reveals that the micro-urokinase molecules pack forming a solvent channel that leads to the active site groove. Therefore, small molecule inhibitors may diffuse into the crystal and bind at the active site. This is important from a structure-based drug design perspective because it facilitates soaking as a method of forming protein-compound complex crystals. The soaking method was used to obtain crystal structures with the three known urokinase inhibitors, B428, amiloride, and phenylguanidine. These structures were obtained at high resolution and provide a starting point for structure-based drug design of a nonpeptidic urokinase inhibitor.

B428—B428 has been reported to inhibit human urokinase with an IC_{50} value of $0.320 \mu M$ (Refs. 25 and 26 and Table II). B428 inhibition was tested *versus* LMW urokinase and micro-urokinase, and Fig. 2 presents the Lineweaver-Burke analysis for the effect of B428 on the activity of micro-urokinase. The results show that B428 competitively inhibits micro-urokinase as observed for the native enzyme (25, 26). As listed in Table II, B428 inhibits LMW urokinase with a K_i of $0.490 \mu M$ while inhibiting micro-urokinase with a K_i of $0.512 \mu M$. Hence, K_i values for the native and re-engineered forms of the protein are essentially identical and are consistent with reported IC_{50} values (25, 26).

The B428-micro-urokinase co-crystal structure was completed to 2.0 \AA resolution. In the complex structure, the $2F_o - F_c$ and $F_o - F_c$ maps indicate that His⁹⁹ is in two conformations as observed in the native structure although Cys⁴² is observed only in the conformation in which the Cys⁴²–Cys⁵⁸ disulfide bridge is intact. It is unclear why only one conformation is observed for the Cys⁴²–Cys⁵⁸ disulfide. In the native structure, the alternate conformation became visible at high resolution. Hence, one possibility is that second conformation is not visible in the lower resolution electron density map. Another explanation is that inhibitor binding may induce a shift to a single conformation or that the inhibitor may only bind to the protein form where the disulfide is intact. Further experiments at high resolution will be necessary to fully understand this phenomenon. Fig. 3A shows the $2F_o - F_c$ (contoured at 1σ) and $F_o - F_c$ (contoured at 3σ) electron density maps calculated in the

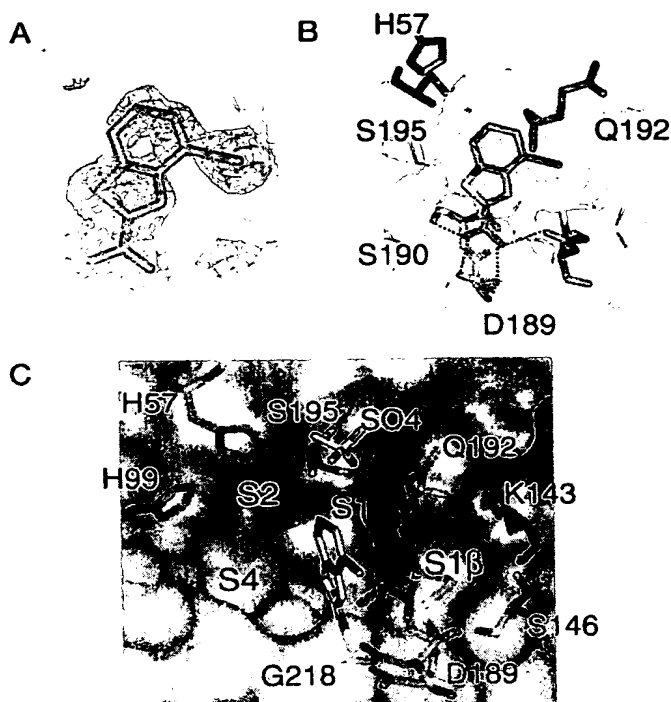


FIG. 3. A, initial $2F_o - F_c$ (purple) and $F_o - F_c$ (green) maps contoured at 1 and 3σ , respectively, for the binding site of B428 before refinement. B, molecular surface as calculated by the program package QUANTA (Molecular Simulations Inc.) depicting interactions between B428 and micro-urokinase. The inhibitor and inhibitor surface are shown in orange, whereas the protein and the protein surface are shown in cyan. C, view of B428 bound at the S_1 site of urokinase. The S_2 site between His⁵⁷ and His⁹⁹ is also shown as well as the S_4 site. An ordered sulfate ion is also shown bound near the oxyanion hole.

absence of inhibitor and before any refinement cycles. All atoms of the inhibitor are clearly defined in both maps, and the compound is found to bind at the S_1 pocket as might be predicted from its net positive charge.

Interactions between B428 and the S_1 pocket are consistent with observations for trypsin and other trypsin-like enzymes (41–45). Nearly all atoms of B428 are in van der Waals' or hydrogen bonding contact with the S_1 site (Fig. 3, B and C). The inhibitor does not occupy other pockets of the substrate-binding groove. The benzothienophene ring is in contact with the rim of the S_1 site that is composed of the Cys¹⁹¹–Cys²²⁰ disulfide bridge and the main chain atoms of Ser²¹⁴–Cys²²⁰ and Gln¹⁹²–Cys¹⁹¹. In the pocket, the thiophene ring is also in contact with the side chains of Val²¹³, Ser¹⁹⁰, Asp¹⁹⁴, and Ser¹⁹⁵. The amidine is donating hydrogen bonds to Ser¹⁹⁰–O γ (3.0 \AA), Asp¹⁸⁹–O $\delta 1$ (2.8 \AA), Asp¹⁸⁹–O $\delta 2$ (2.8 \AA), and Gly²¹⁸–O (2.7 \AA) (Fig. 3B). Hence, both hydrophobic and hydrophilic interactions occur at S_1 .

In addition to interactions at S_1 , the 4-iodo group is pointing out of the S_1 pocket away from the substrate-binding groove and is making van der Waals' interactions with the side chain of Cys²²⁰ and the main chain atoms of Gly²¹⁸. These residues form part of a subpocket composed of the disulfide bridge at Cys¹⁹¹–Cys²²⁰, residues Gly²¹⁸ and Ser¹⁴⁶, and the side chain of Lys¹⁴³. This pocket has been termed the $S_{1\beta}$ pocket because of its proximity to the primary S_1 site (Fig. 3C). It is reported that the 4-iodo group of B428 confers a 10-fold increase in binding potency relative to the 4-hydro compound (25, 26). This observation is consistent with the B428-urokinase crystal structure where the 4-iodo group partially accesses the $S_{1\beta}$ pocket. Fur-

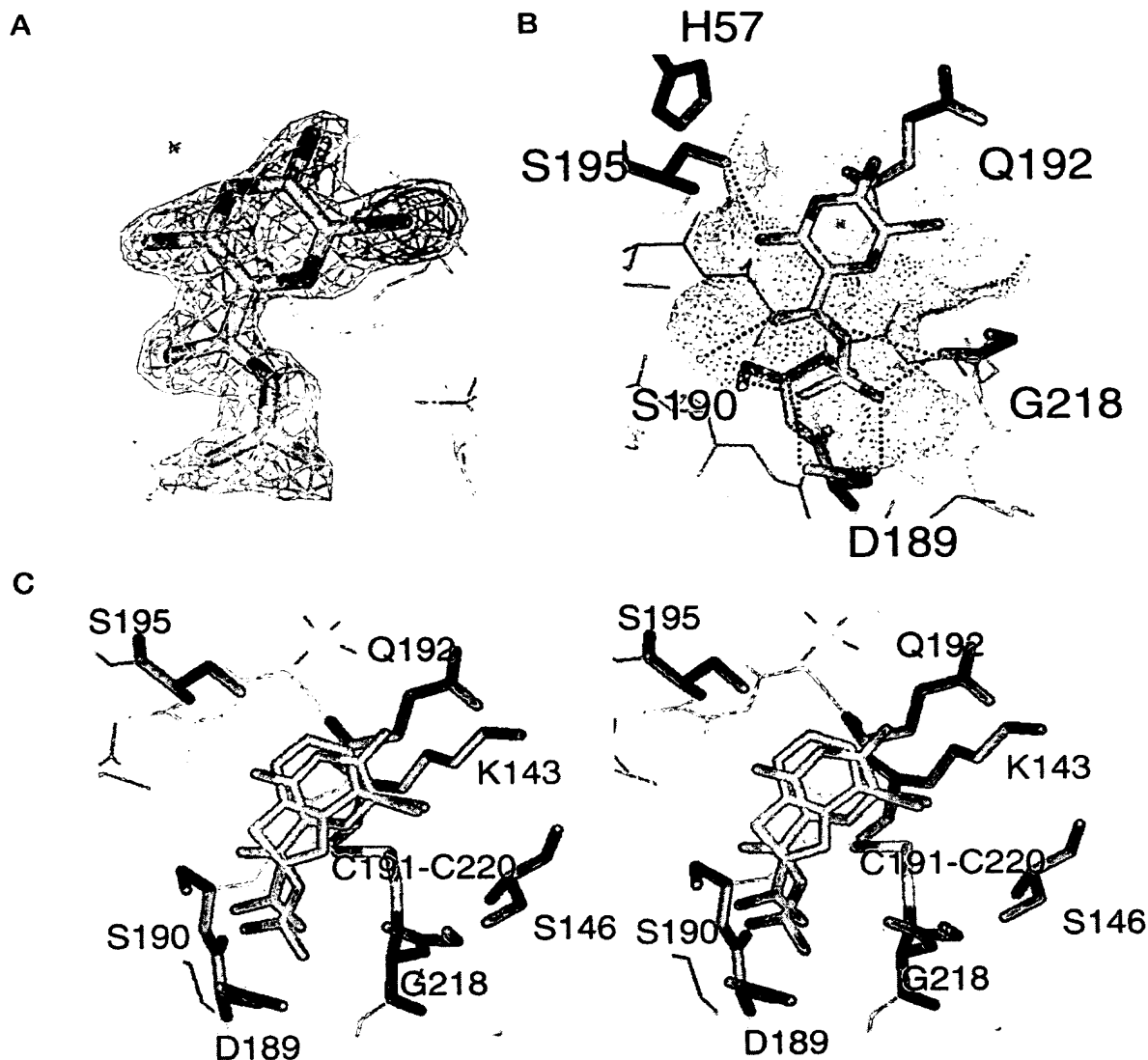


FIG. 4. *A*, initial $2F_o - F_c$ (purple) and $F_o - F_c$ (green) maps contoured at 1 and 3 σ , respectively, for the binding site of amiloride before refinement. *B*, molecular surface as calculated by the program package QUANTA (Molecular Simulations Inc.) depicting interactions between amiloride and micro-urokinase. The inhibitor and inhibitor surface are shown in peach, whereas the protein and protein surface are shown in cyan. *C*, overlay of the crystal structures of amiloride (purple) and B428 (orange) micro-urokinase showing that the halogen atoms of each inhibitor are occupying the same site.

thermore, B623 inhibits urokinase with an IC_{50} of $0.07 \mu M$ (25, 26). Based upon the crystal structure of B428-micro-urokinase, it is possible that this larger 4-substituent is occupying more of the $S_1\beta$ pocket³ and consequently binds more tightly to urokinase. Hence, access to this novel pocket has been shown to confer an increase in binding potency and may serve as a site for further substitution in structure-based drug design.

Examination of the crystal structure of B428-urokinase shows that the 5 and 6 positions of the benzo(b)thiophene-2-carboxamide are also open for substitution, whereas the 3 and 7 positions are buried within the S_1 pocket and therefore less likely to accommodate a substituent. Of these, the 5 position does not directly point toward any pockets of the urokinase molecule because it points toward Gln¹⁹² and out toward bulk solvent. Hence, substitution at this position is less likely to

confer a large increase in binding potency. On the other hand, the 6 position points toward the urokinase catalytic site although the position appears partially blocked by the side chain of the active site Ser¹⁹⁵. The distance from Ser¹⁹⁵-OH to the 6 position carbon is 3.2 Å; therefore incorporation of a substitution at this position may require a shifting of the benzothio-phenene scaffold away from Ser¹⁹⁵. Additionally, substitutions at the 6 position would not orient toward the substrate-binding groove accessed by Glu-Gly-Arg-chloromethyl ketone. Substitutions at the 6 position would have to bend back toward the substrate-binding site or access other subsites. Nevertheless, the 4 and 6 positions appear to be the best substitution sites toward increasing the binding potency of B428, and both sets of substitutions will likely occupy sites apart from the substrate-binding groove.

Amiloride—Amiloride has been reported to inhibit human urokinase with a K_i (24) or IC_{50} of $7 \mu M$ (25, 26). As observed with B428, amiloride also competitively inhibits LMW uroki-

³ The crystal structure of B623 in complex with urokinase could not be completed because of solubility issues with the compound.

nase and micro-urokinase with similar values ($K_i = 7.2 \mu\text{M}$ for LMW urokinase, and $K_i = 6.9 \mu\text{M}$ for micro-urokinase). Amiloride is a weaker urokinase inhibitor than B428 (Table II) but may have more favorable pharmacological properties because the compound is an orally active commercial drug (46). To compare the binding modes of amiloride and B428 and to establish strategies for development of a more potent amiloride-based urokinase inhibitor, the co-crystal structure of amiloride micro-urokinase was completed at 2.2 Å resolution.

Examination of the $2F_o - F_c$ (contoured at 1 σ) and $F_o - F_c$ (contoured at 3 σ) electron density maps at the active site shows that all atoms of the inhibitor are clearly defined in both maps (Fig. 4A). In addition, the maps show His⁹⁹ in two conformations and the Cys⁴²-Cys⁵⁸ disulfide bridge intact as observed in the B428 complex. The data also indicate that amiloride binds at the S₁ pocket as observed with B428 (Fig. 4C).

The crystal structure of amiloride-micro-urokinase indicates that amiloride is making more hydrogen bonding interactions at the S₁ site than B428 while maintaining some of the van der Waals' interactions within the pocket. The size of the amiloride pyrazine scaffold is smaller than the B428 benzothiophene such that even though the pyrazine ring is in contact with the rim of the S₁ pocket as observed for B428, the extent of the packing interactions is smaller. In place of the thiophene ring, the 3-amino and 2-acylguanidine groups of amiloride are making hydrogen bonding interactions. Specifically, the 3-amino group is packed underneath the side chain of Ser¹⁹⁵ as shown in Fig. 4B where it is donating a hydrogen bond to Ser¹⁹⁵-O γ (3.1 Å). The carbonyl of the acyl guanidine group is accepting a hydrogen bond (2.9 Å) from a buried solvent molecule bound directly above Tyr²²⁸. The guanidine-NH is donating a hydrogen bond to Gly²¹⁸-O (3.1 Å). As observed with B428, the amide-like nitrogens are donating hydrogen bonds to Gly²¹⁸-O (2.7 Å) and Asp¹⁸⁹-O δ 1 (3.0 Å) or to Asp¹⁸⁹-O δ 2 (3.0 Å), and Ser¹⁹⁰-O γ (2.7 Å). The hydrogen bonding geometry of the guanidinium group is also very similar to that observed for ArgP₁ in the Glu-Gly-Arg-chloromethyl ketone-LMW urokinase structure (1). Hence, although the core scaffolds of both B428 and amiloride are bound at the S₁ pocket, the nature of the interactions within the pocket are different.

The crystal structure of amiloride-micro-urokinase reveals strategies for structure-based drug design of a more potent small molecule inhibitor. One potential site of substitution is the 6 position. The 6-chloro group of amiloride is accessing the S₁ β pocket as observed for the 4-iodo group of B428. Specifically the 6-chloro group is in hydrophobic contact with the side chain of Cys²²⁰ and the main chain atoms of Gly²¹⁸ (Fig. 4C). Thus, although the chemical structures of B428 and amiloride are very different, interactions at the S₁ β pocket are nearly identical. Because of this similarity, one might substitute the 6-chloro position of amiloride with larger groups such as iodine (present in B428) or a benzodioxol arylethenyl (present in B623), which were both shown to enhance the activity in the benzo(b)thiophene-2-carboxamidine series. The 3 position of amiloride within the S₁ pocket is another site for substitution. However, substitutions at this site are expected to point toward Gln¹⁹² and then out toward bulk solvent as observed for the 5 position of B428. Thus, use of a rigid linker may be necessary to redirect substitutions toward the protein including the substrate-binding groove. In summary, substitutions of the amiloride scaffold should occur at the 5 and 6 positions to provide direct access to the S₁ β pocket or indirect access to other sites on the protein.

Phenylguanidine—Phenylguanidine inhibits urokinase with a K_i of 20.6 μM (27) and is therefore a weaker inhibitor of urokinase than either amiloride or B428 (Table II). This inhib-

itor also competitively inhibits micro-urokinase with a K_i consistent with the LMW form ($K_i = 20.6 \mu\text{M}$ LMW for urokinase, and $K_i = 17.4 \mu\text{M}$ for micro-urokinase). To compare the binding mode of this inhibitor to amiloride and B428 and to determine potential sites of substitution, the co-crystal structure of phenylguanidine-micro-urokinase was completed at 2.0 Å resolution.

The phenylguanidine-micro-urokinase active site structure is very similar to that in the presence of B428 and amiloride. His⁹⁹ is observed in multiple conformations while the Cys⁴²-Cys⁵⁸ disulfide bridge is intact. Additionally, the $2F_o - F_c$ (contoured at 1 σ) and $F_o - F_c$ (contoured at 3 σ) electron density maps (Fig. 5A) obtained using the urokinase model in the absence of inhibitor and before any refinement cycles shows that all atoms of the inhibitor are clearly defined in both maps. The inhibitor was found to bind at the S₁ pocket (Fig. 5B).

Even though both amiloride and phenylguanidine have scaffolds of the same size, the phenyl ring of phenylguanidine binds very differently from the pyrazine ring of amiloride (Fig. 5, B and C). Specifically, the phenylguanidine ring packs underneath Ser¹⁹⁵ and is interacting with the main chain atoms of Val²¹³-Trp²¹⁵ as well as the side chain of Val²¹³. The ring also interacts with the main chain atoms of Ser¹⁹⁰-Cys¹⁹¹ as well as the side chain of Ser¹⁹⁰. The differential ring packing is most likely due to amiloride possessing one additional linker atom between the guanidine and aromatic groups relative to phenylguanidine (Table II) because the guanidine groups are oriented very similarly. Specifically, the guanidine-NH is donating a hydrogen bond to Gly²¹⁸-O (3.0 Å), whereas the amidine-like nitrogens are donating hydrogen bonds to Gly²¹⁸-O (2.9 Å) and Asp¹⁸⁹-O δ 1 (2.9 Å) or to Asp¹⁸⁹-O δ 2 (3.0 Å) and Ser¹⁹⁰-O γ (3.3 Å). Thus, it is likely that the core scaffold of amiloride (pyrazine ring) orients differently than the phenyl group of phenylguanidine because the binding is being driven by the hydrogen bonding geometry of the guanidine groups rather than the van der Waals/hydrogen bonding interactions of the core groups even though interactions of the core groups most certainly contribute to the compound binding.

The phenyl guanidine urokinase structure also shows that Gln¹⁹² has changed conformation and is in hydrophobic contact with the inhibitor (Fig. 5B) such that it is blocking the entrance to the S₁ β pocket. In the native and the B428 or amiloride complex structures, the S₁ β pocket is open where Gln¹⁹² is accepting a hydrogen bond from Lys¹⁴³ (3.3 Å) and donating a hydrogen bond to Tyr¹⁵¹ (3.1 Å). Thus, a conformational shift of this side chain requires breaking two hydrogen bonds. This is not the case for other serine proteases such as thrombin where there is no hydrogen bonding partner for Glu¹⁹² in either position. Here, there is less of an energy barrier to a conformational shift of Glu¹⁹², and the side chain may be found in both conformations (49, 50). For urokinase, it appears that the binding of certain inhibitors such as phenyl guanidine does break the two Gln¹⁹² hydrogen bonds and conformationally shift Gln¹⁹² to maximize hydrophobic desolvation of the compound. Hence, Gln¹⁹² may be induced to shift conformation and because Gln¹⁹² may act as a switch to the entrance to S₁ β from S₁, noting the orientation of this side chain is important in a drug design strategy.

The crystal structure of phenylguanidine-urokinase suggests a structure-based drug design strategy different from that with B428 or amiloride. Both B428 and amiloride are capable of directly accessing the S₁ β pocket, whereas the binding orientation of phenylguanidine is such that a similar interaction cannot be achieved by direct substitution of the phenyl ring (Fig. 5C) even with movement of Gln¹⁹² to the S₁ β open position. Specifically, as shown in Fig. 5 (B and C), the 2 and 3

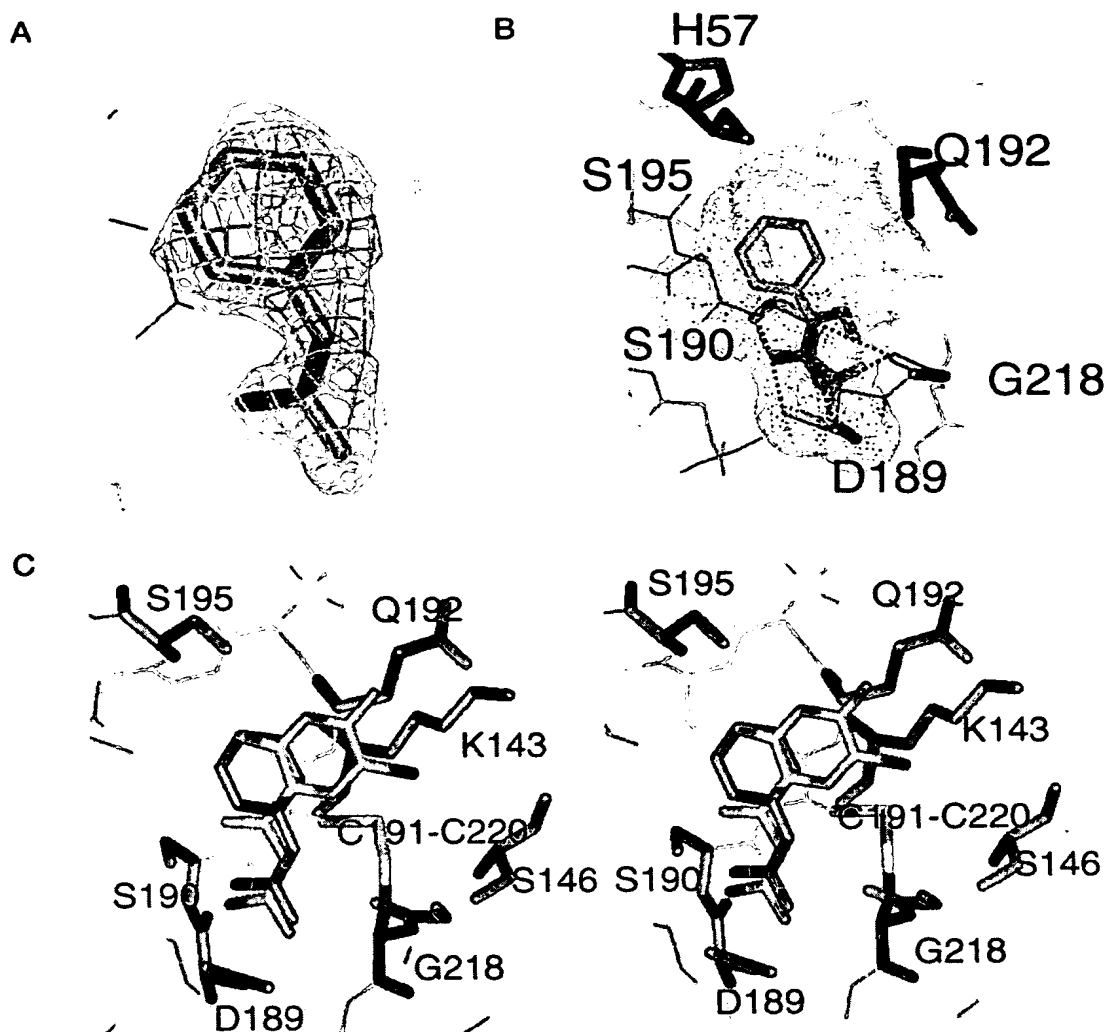


FIG. 5. A, initial $2F_o - F_c$ (purple) and $F_o - F_c$ (green) maps contoured at 1 and 3 σ , respectively, for the binding site of phenyl guanidine before refinement. B, molecular surface micro-urokinase as calculated by the program package QUANTA (Molecular Simulations Inc.) depicting interactions between B428 and micro-urokinase. The inhibitor and inhibitor surface are shown in orange, whereas the protein and protein surface are shown in cyan. C, overlay of the crystal structures of amiloride (purple) and phenyl guanidine (black) micro-urokinase, showing that the two scaffolds occupy different areas of the S_1 pocket.

positions could point toward the $S_{1\beta}$ pocket but are too far away to support direct interaction with $S_{1\beta}$. In fact, substitution of the phenyl ring with halogens at both the 2 and 3 positions did not result in any increase in inhibitory potency (27). On the other hand, substitution at position 4 with a chloro- or trifluoromethyl-group resulted in an increase in inhibition to K_i values of 6.8 and 6.5 μM , respectively (27). This 4 substitution is expected to orient toward the side chain of Ser¹⁹⁵ and may obtain binding energy from a favorable van der Waals' packing interaction with Ser¹⁹⁵ and the S_1 pocket. The 5 and 6 positions are within the S_1 pocket and therefore less open for substitution. Because interactions with the $S_{1\beta}$ pocket are expected to confer an increase in binding potency and because phenylguanidine may not directly access this site, modification of the scaffold may be a promising drug design strategy for this series.

Further examination of an overlay of the crystal structures of phenyl guanidine and amiloride micro-urokinase (Fig. 5C) shows that the binding of the two scaffolds is complementary. The lack of overlap between the two groups suggests that the phenyl and pyrazine rings could be fused to form a 1-naphthyl-

ylguanidine system. The naphthyl ring would be expected to occupy the sites of both core scaffolds and could therefore maintain the positive characteristics of both the phenylguanidine and amiloride series. This would include utilization of the 4-chloro or 4-trifluoromethyl substitutions in the phenylguanidine series as well as access to the $S_{1\beta}$ pocket exploited by amiloride and B428. Hence, a merging of the amiloride and phenylguanidine scaffolds would be predicted to benefit from the additivity of both sites and create a more potent and easily optimized urokinase inhibitor.

DISCUSSION

Urokinase inhibitors have been shown to affect tumor metastasis and growth *in vivo* making urokinase an attractive anti-cancer target. However, these existing compounds lack all of the properties necessary for a therapeutic agent and require optimization. Crystallography driven structure-based drug design based on a series of ligand-protein crystal structures can be utilized to optimize urokinase inhibition. The properties of the protein crystals can affect the efficiency of structure-based drug design because a larger number of more accurate struc-

tures provides a better description of the relationship between binding interactions and binding energy. Fortunately, advances in molecular biology can be used to engineer the protein to obtain crystal systems that facilitate faster and more exact structure determinations and enhance the drug design cycle (47). Such a method has been used to design a crystal system for human urokinase for optimization of a urokinase inhibitor.

The sequence of LMW urokinase was redesigned to produce a new crystal form that would permit a more ideal system for structure-based drug design. Specifically, LMW urokinase was re-engineered to minimize the areas of disorder that may likely cause suboptimal crystal packing. This recombinant protein, micro-urokinase, produces crystals with close packing interactions at the A-chain cleft, which would be blocked in LMW urokinase. This close molecular packing results in crystals that diffract to high resolution on a rotating anode source (1.6–2.0 Å). However, even though the micro-urokinase molecules are closely packed, the active site is both unoccupied and open to solvent channels in the crystal. This property readily allows compounds to be diffused into the crystal and has facilitated the determination of crystal structures in the presence of three reported urokinase inhibitors toward design of an anti-cancer agent.

The micro-urokinase crystal system and soaking method was used to determine the co-crystal structures of micro-urokinase complexed with the inhibitors B428 (25, 26), amiloride (24), and phenylguanidine (27). Each of the co-crystal structures gives insight into favorable compound-protein interactions that contribute to the binding of these inhibitors to urokinase. The primary binding force is likely the hydrogen bonds between each inhibitor's amidine or guanidine group and Asp¹⁸⁹. This salt bridge interaction is common to many guanidine or amidine complexes with trypsin or trypsin-like serine proteases such as thrombin, factor Xa, or tissue plasminogen activator (41–45) and is observed for Arg-P₁ in the Glu-Gly-Arg-chloromethyl ketone LMW urokinase structure (1). In addition to the hydrogen bonding interactions, van der Waals' packing between the core scaffold and the S₁ pocket may also contribute to the overall binding energy. Hydrophobic packing at the S₁ pocket is the primary binding interaction between substrates/inhibitors in the chymotrypsin family of proteases where the S₁ pocket contains no charged groups (48–51). Additionally, a series of thrombin inhibitors that lack a positively charged group to interact with Asp¹⁸⁹ have been described (52, 53). Hence, both hydrophilic and hydrophobic interactions at the S₁ pocket contribute to the binding of B428, amiloride, and phenylguanidine, and these interactions are present in other crystal structures.

Examination of the urokinase structures reveals a new additional binding site adjacent to the S₁ pocket. The site, termed the S₁β subpocket, is composed of the disulfide bridge at Cys¹⁹¹–Cys²²⁰, residues Ser¹⁴⁶ and Gly²¹⁸, and the side chain of Lys²¹⁴. The S₁β subpocket is also present in the LMW urokinase structure (Protein Data Bank entry 1LMW) and is away from any re-engineered sites. The crystal structure of phenylguanidine urokinase reveals that Gln¹⁹² may act as a switch for the closing and opening of S₁β. In the native and B428 or amiloride complex structures, the S₁β pocket is open, and Gln¹⁹² is involved in two hydrogen bonds (Lys¹⁴³ and Tyr¹⁵¹). However, in the presence of other inhibitors such as phenylguanidine or Glu-Gly-Arg-chloromethyl ketone (1), the hydrogen bonds are broken, and the conformation of Gln¹⁹² shifted such that its side chain is in van der Waals' contact with the inhibitor. In this conformation, the entrance to S₁β is blocked, and the shift is most likely induced to maximize interactions with the inhibitor. Hence, although the S₁β pocket may be

blocked by the induced movement of Gln¹⁹², its proximity to S₁ makes it an attractive subsite for structure-based drug design.

The halogen atoms of B428 and amiloride are interacting with the entrance to the S₁β subsite (Gly²¹⁸–Cys²²⁰). Interactions at this site have been shown to confer a significant increase in inhibitory potency for the benzo(b)thiophene-2-carboxamide series where the 4-iodo group (IC₅₀ = 0.32 μM) or 4-benzodioxolanylethyl (IC₅₀ = 0.07 μM) inhibit more strongly than the 4-hydro compound (IC₅₀ = 3.7 μM) (25, 26). The increase in potency observed for both substitutions is most likely due to packing interactions at the S₁β pocket. Phenylguanidine lacks a halogen atom to access the S₁β pocket, and examination of the structure reveals that the pocket can not be easily accessed by a direct substitution of the phenylguanidine ring. However, an overlay of the phenylguanidine crystal structure with that of amiloride reveals that the two scaffolds could be merged to form a 1-guanadyl naphthalene. This compound could, in turn, access the S₁β pocket. Hence, urokinase co-crystal structures with B428, amiloride, and phenylguanidine indicate that all three scaffolds may provide either direct or indirect access to the S₁β pocket. Furthermore, this newly described subsite has great potential for the future design of more potent urokinase inhibitors for the treatment of cancer.

Acknowledgments—We thank Dr. Bruce Littlefield of the Eisai Company for initial supplies of B428 and Dr. Todd Rockway for synthesis of ϵ -amino caproic acid *p*-carboxyphenyl ester chloride. We also thank Dr. Stephen Betz for critical examination of the manuscript and Dr. Jonathan Greer for many helpful discussions and critical examination of the manuscript.

REFERENCES

1. Spraggon, G., Phillips, C., Nowak, U. K., Ponting, C. P., Saunders, D., and Dobson, C. M. (1995) *Structure* **3**, 681–691.
2. Kohn, E. C. (1991) *Pharmacol. Ther.* **52**, 235–244.
3. Quax, P. H., van, L. R. T., Verspaget, H. W., and Verheijen, J. H. (1990) *Cancer Res.* **50**, 1488–1494.
4. Behrendt, N., Ronne, E., Ploug, M., Petri, T., Lober, D., Nielsen, L. S., Schleuning, W. D., Blasi, F., Appella, E., and Dano, K. (1990) *J. Biol. Chem.* **265**, 6453–6460.
5. Schmitt, M., Janicke, F., Moniwa, N., Chucholowski, N., and Pache. (1992) *Biol. Chem. Hoppe-Seyler* **373**, 611–627.
6. Duffy, M. J. (1990) *Blood Coagul. Fibrinolysis* **1**, 681–687.
7. Astedt, B., Billstrom, A., and Lécander, I. (1995) *Fibrinolysis* **9**, 175–177.
8. Evans, D., Sloan-Stakleff, K., Arvan, M., and Guyton, D. (1998) *Clin. Exp. Metastasis* **16**, 353–357.
9. Banerji, A., Fernandes, A., Bane, S., and Ahire, S. (1998) *Cancer Lett.* **129**, 15–20.
10. Kobayashi, H., Gotoh, J., Shinohara, H., Moniwa, N., and Terao, T. (1994) *Thromb. Haemostasis* **71**, 474–480.
11. Xiao, G., Liu, Y., Gentz, R., Sang, Q., Goldberg, I., and Shi, Y. (1999) *Proc. Nat. Acad. Sci. U. S. A.* **96**, 3700–3705.
12. Rabbani, S., Harakidas, P., Davidson, D., Henkin, J., and Mazar, A. (1995) *Int. J. Cancer* **63**, 840–845.
13. Alonso, D., Tejera, A., Farias, E., Joffe, E., and Bomez, D. (1998) *Anticancer Res.* **18**, 4499–4504.
14. Alonso, D., Farias, E., Ladeda, V., Davel, L., Puricelli, L., and Joffe, E. (1996) *Breast Cancer Res. Treat.* **40**, 209–223.
15. Jankun, J., Keck, R., Skrzypczak-Jankun, E., and Swiercz, R. (1997) *Cancer Res.* **57**, 559–563.
16. Browner, M. F., Smith, W. W., and Castelano, A. L. (1995) *Biochemistry* **34**, 6602–6610.
17. Chand, P., Babu, Y. S., Bantia, S., Chu, N., Cole, L. B., and Kotian, P. L. (1997) *J. Med. Chem.* **40**, 4030–4052.
18. Erickson, J., Neidhart, D. J., VanDrie, J., Kempf, D. J., and Wang, X. C. (1990) *Science* **249**, 527–533.
19. Lam, P. Y., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., and Ru, Y. (1994) *Science* **263**, 380–384.
20. Kurumbail, R. G., Stevens, A. M., Gierse, J. K., McDonald, J. J., Stegeman, R. A., and Pak, J. Y. (1996) *Nature* **384**, 644–648.
21. Luong, C., Miller, A., Barnett, J., Chow, J., and Ramesha, C. (1996) *Nat. Struct. Biol.* **3**, 927–933.
22. Verlinde, C. L., and Hol, W. G. (1994) *Structure* **2**, 577–587.
23. Luzatti, P. V. (1952) *Acta Crystallogr.* **5**, 802–810.
24. Vassalli, J. D., and Belin, D. (1987) *FEBS Lett.* **214**, 187–191.
25. Bridges, A. J., Lee, A., Schwartz, C. E., Towle, M. J., and Littlefield, B. A. (1993) *Bioorg. Med. Chem.* **1**, 403–410.
26. Towle, M. J., Lee, A., Maduakor, E. C., Schwartz, C. E., Bridges, A. J., and Littlefield, B. A. (1993) *Cancer Res.* **53**, 2553–2559.
27. Yang, H., Henkin, J., Kim, K. H., and Greer, J. (1990) *J. Med. Chem.* **33**, 2956–2961.
28. Lo, K.-M., and Gillies, S. D. (1991) *Biochi. Biophys. Acta* **1088**, 217–224.
29. Wang, J., Brdar, B., and Reich, E. (1995) *Protein Sci.* **4**, 1758–1767.

30. Barlow, G. H. (1976) *Methods Enzymol.* **45**, 239–244
31. Segel, I. H. (1975) *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*, John Wiley & Sons, New York
32. Menegatti, E., Guarneri, M., Bolognesi, M., Ascenzi, P., and G., A. (1989) *J. Enzyme Inhibition* **2**, 249–259
33. Otwinowski, Z., and Minor, W. (1997) *Methods Enzymol.* **276**, 307–326
34. Navaza, J. (1994) *Acta Crystallogr. A* **50**, 157–163
35. Brunger, A. T. (1993) *X-PLOR*, version 3.1, Yale University Press, New Haven, CT
36. Sheldrick, G. M. (1990) *SHELX-97*, Gottingen University, Gottingen, Germany
37. Bode, W., Mayr, I., Baumann, U., Huber, R., and Stone, S. R. (1989) *EMBO J.* **8**, 3467–3475
38. Lamba, D., Bauer, M., Huber, R., Fischer, S., Rudolph, R., Kohnert, U., and Bode, W. (1996) *J. Mol. Biol.* **258**, 117–135
39. Padmanabhan, K., Padmanabhan, K. P., Tulinsky, A., Park, C. H., Bode, W., Huber, R., Blankenship, D. T., Cardin, A. D., and Kisiel, W. (1993) *J. Mol. Biol.* **232**, 947–966
40. Henderson, R. (1970) *J. Mol. Biol.* **54**, 341–354
41. Bode, Q., Turk, D., and Sturzebecher, J. (1990) *Eur. J. Biochem.* **193**, 175–182
42. Bode, W., and Schwager, P. (1975) *J. Mol. Biol.* **98**, 693–717
43. Banner, D. W., and Hadvary, P. (1991) *J. Biol. Chem.* **266**, 20085–20093
44. Renatus, M., Bode, W., Huber, R., Sturzebecher, J., Prasa, D., Fischer, S., Kohnert, U., and Stubbs, M. (1997) *J. Biol. Chem.* **272**, 21713–21719
45. Brandstetter, H., Kuhne, A., Bode, W., Huber, R., von der Saal, W., Wirthensohn, K., and Engh, R. (1996) *J. Biol. Chem.* **271**, 29988–29992
46. Baba, W. I., Lant, A. F., Smith, A. J., Townshend, M. M., and Wilson, G. M. (1968) *Clin. Pharmacol. Ther.* **9**, 318–327
47. Price, S., and Nagai, K. (1995) *Curr. Opin. Biotechnol.* **6**, 425–430
48. Steitz, T., Henderson, R., and Blow, D. (1969) *J. Mol. Biol.* **46**, 337–348
49. Blow, D. (1976) *Acc. Chem. Res.* **9**, 145–152
50. Sigler, P., Jeffery, B., Matthews, B., and Blow, D. (1966) *J. Mol. Biol.* **15**, 175–192
51. Birktoft, J., and Blow, D. (1972) *J. Mol. Biol.* **68**, 187–240
52. Malikayil, J. A., Burkhart, J. P., Schreuder, H. A., Broersma, R. J., Tardif, C., Kutcher, L. W., Mehdi, S., Schatzman, G. L., Neises, B., and Peet, N. P. (1997) *Biochemistry* **36**, 1034–1040
53. Das, J., and Kimball, S. D. (1995) *Bioorg. Med. Chem.* **3**, 999–1007

Exhibit 25

PCTWORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : C07K 14/435, 14/705, A61K 38/03, 38/08, 38/17	A1	(11) International Publication Number: WO 00/52044 (43) International Publication Date: 8 September 2000 (08.09.00)
(21) International Application Number: PCT/US00/05612 (22) International Filing Date: 2 March 2000 (02.03.00) (30) Priority Data: 09/261,416 3 March 1999 (03.03.99) US (71) Applicant: THE BOARD OF TRUSTEES OF THE UNIVERSITY OF ARKANSAS [US/US]; 2404 North University Avenue, Little Rock, AR 72207-3608 (US). (72) Inventors: O'BRIEN, Timothy, J.; 2610 North Pierce, Little Rock, AR 72207 (US). UNDERWOOD, Lowell, J.; Apartment K, 121 N. Jackson Street, Little Rock, AR 72205 (US). (74) Agent: ADLER, Benjamin, A.; McGregor & Adler, 8011 Candle Lane, Houston, TX 77071 (US).		(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>With international search report.</i> <i>Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>
(54) Title: TRANSMEMBRANE SERINE PROTEASE OVEREXPRESSED IN OVARIAN CARCINOMA AND USES THEREOF (57) Abstract The present invention provides a TADG-12 protein and a DNA fragment encoding such protein. Also provided is a vector/host cell capable of expressing the DNA. The present invention further provided various methods of early detection of associated ovarian and other malignancies, and of interactive therapies for cancer treatment by utilizing the DNA and/or protein disclosed herein.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakistan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

**TRANSMEMBRANE SERINE PROTEASE OVEREXPRESSED IN
OVARIAN CARCINOMA AND USES THEREOF**

BACKGROUND OF THE INVENTION

Cross-Reference to Related Application

This application is a continuation-in-part patent application and claims the benefit of priority under 35 USC §120 of USSN 09/261,416, filed March 3, 1999.

Field of the Invention

The present invention relates generally to the fields of cellular biology and diagnosis of neoplastic disease. More specifically, the present invention relates to a transmembrane serine protease termed Tumor Associated Differentially-Expressed Gene-12 (TADG-12), which is overexpressed in ovarian carcinoma.

Description of the Related Art

Tumor cells rely on the expression of a concert of proteases to be released from their primary sites and move to distant sites to inflict lethality. This metastatic nature is the result of an aberrant expression pattern of proteases by tumor cells and also by stromal cells surrounding the tumors [1-3]. For most tumors to become metastatic, they must degrade their surrounding extracellular matrix components, degrade basement

membranes to gain access to the bloodstream or lymph system, and repeat this process in reverse fashion to settle in a secondary host site [3-6]. All of these processes rely upon what now appears to be a synchronized protease cascade. In addition, tumor cells
5 use the power of proteases to activate growth and angiogenic factors that allow the tumor to grow progressively [1]. Therefore, much research has been aimed at the identification of tumor-associated proteases and the inhibition of these enzymes for therapeutic means. More importantly, the secreted nature and/or
10 high level expression of many of these proteases allows for their detection at aberrant levels in patient serum, e.g. the prostate-specific antigen (PSA), which allows for early diagnosis of prostate cancer [7].

Proteases have been associated directly with tumor
15 growth, shedding of tumor cells and invasion of target organs. Individual classes of proteases are involved in, but not limited to (1) the digestion of stroma surrounding the initial tumor area, (2) the digestion of the cellular adhesion molecules to allow dissociation of tumor cells; and (3) the invasion of the basement
20 membrane for metastatic growth and the activation of both tumor growth factors and angiogenic factors.

For many forms of cancer, diagnosis and treatment has improved dramatically in the last 10 years. However, the five year survival rate for ovarian cancer remains below 50% due in
25 large part to the vague symptoms which allow for progression of the disease to an advanced stage prior to diagnosis [8]. Although the exploitation of the CA125 antigen has been useful as a marker for monitoring recurrence of ovarian cancer, it has not proven to be an ideal marker for early diagnosis. Therefore, new markers

that may be secreted or released from cells and which are highly expressed by ovarian tumors could provide a useful tool for the early diagnosis and for therapeutic intervention in patients with ovarian carcinoma.

5 The prior art is deficient in the lack of the complete identification of the proteases overexpressed in carcinoma, therefore, deficient in the lack of a tumor marker useful as an indicator of early disease, particularly for ovarian cancers. Specifically, TADG-12, a transmembrane serine protease, has not
10 been previously identified in either nucleic acid or protein form. The present invention fulfills this long-standing need and desire in the art.

SUMMARY OF THE INVENTION

15

 The present invention discloses TADG-12, a new member of the Tumor Associated Differentially-Expressed Gene (TADG) family, and a variant splicing form of TADG-12 (TADG-12V) that could lead to a truncated protein product. TADG-12 is a
20 transmembrane serine protease overexpressed in ovarian carcinoma. The entire cDNA of TADG-12 has been identified (SEQ ID No. 1). This sequence encodes a putative protein of 454 amino acids (SEQ ID No. 2) which includes a potential transmembrane domain, an LDL receptor like domain, a scavenger receptor
25 cysteine rich domain, and a serine protease domain. These features imply that TADG-12 is expressed at the cell surface, and it may be used as a molecular target for therapy or a diagnostic marker.

In one embodiment of the present invention, there is provided a DNA fragment encoding a TADG-12 protein selected from the group consisting of: (a) an isolated DNA fragment which encodes a TADG-12 protein; (b) an isolated DNA fragment which hybridizes to isolated DNA fragment of (a) above and which encodes a TADG-12 protein; and (c) an isolated DNA fragment differing from the isolated DNA fragments of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-12 protein. Specifically, the DNA fragment has a sequence shown in SEQ ID No. 1 or SEQ ID No. 3.

In another embodiment of the present invention, there is provided a vector/host cell capable of expressing the DNA of the present invention.

In yet another embodiment of the present invention, there is provided an isolated and purified TADG-12 protein encoded by DNA selected from the group consisting of: (a) isolated DNA which encodes a TADG-12 protein; (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-12 protein; and (c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-12 protein. Specifically, the TADG-12 protein has an amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4.

In still yet another embodiment of the present invention, there is provided a method for detecting expression of a TADG-12 protein, comprising the steps of: (a) contacting mRNA obtained from the cell with the labeled hybridization probe; and (b) detecting hybridization of the probe with the mRNA.

The present invention further provides methods for diagnosing a cancer or other malignant hyperplasia by detecting the TADG-12 protein or mRNA disclosed herein.

5 In still another embodiment of the present invention, there is provided a method of inhibiting expression of endogenous TADG-12 mRNA in a cell by introducing a vector into the cell, wherein the vector comprises a DNA fragment of TADG-12 in opposite orientation operably linked to elements necessary for expression.

10 In still yet another embodiment of the present invention, there is provided a method of inhibiting expression of a TADG-12 protein in a cell by introducing an antibody directed against a TADG-12 protein or fragment thereof.

15 In still yet another embodiment of the present invention, there is provided a method of targeted therapy by administering a compound having a targeting moiety specific for a TADG-12 protein and a therapeutic moiety. Specifically, the TADG-12 protein has an amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4.

20 The present invention still further provides a method of vaccinating an individual against TADG-12 by inoculating the individual with a TADG-12 protein or fragment thereof. Specifically, the TADG-12 protein has an amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4. The TADG-12 fragment
25 includes the truncated form of TADG-12V peptide having a sequence shown in SEQ ID No. 8, and a 9-residue up to 12-residue fragment of TADG-12 protein.

In yet another embodiment of the present invention, there is provided an immunogenic composition, comprising an

immunogenic fragment of a TADG-12 protein and an appropriate adjuvant. The TADG-12 fragment includes the truncated form of TADG-12V peptide having a sequence shown in SEQ ID No. 8, and a 9-residue up to 12-residue fragment of TADG-12 protein.

5. Other and further aspects, features, and advantages of the present invention will be apparent from the following description of the presently preferred embodiments of the invention given for the purpose of disclosure.

10 BRIEF DESCRIPTION OF THE DRAWINGS

So that the matter in which the above-recited features, advantages and objects of the invention, as well as others which will become clear, are attained and can be understood in detail,
15 more particular descriptions of the invention briefly summarized above may be had by reference to certain embodiments thereof which are illustrated in the appended drawings. These drawings form a part of the specification. It is to be noted, however, that the appended drawings illustrate preferred embodiments of the
20 invention and therefore are not to be considered limiting in their scope.

Figure 1A shows that the expected PCR product of approximately 180 bp and the unexpected PCR product of approximately 300 bp using the redundant serine protease
25 primers were not amplified from normal ovary cDNA (Lane 1) but were found in abundance from ovarian tumor cDNA (Lane 2). The primer sequences for the PCR reactions are indicated by horizontal arrows. **Figure 1B** shows that TADG-12 was subcloned from the 180 bp band while the larger 300 bp band was designated TADG-

12V. The sequences were found to overlap for 180 bp (SEQ ID No. 5 for nucleotide sequence, SEQ ID No. 6 for deduced amino acid sequence) with the 300 bp TADG-12V (SEQ ID No. 7 for nucleotide sequence, SEQ ID No. 8 for deduced amino acid sequence) having an additional insert of 133 bases. This insertion (vertical arrow) leads to a frame shift, which causes the TADG-12V transcript to potentially produce a truncated form of TADG-12 with a variant amino acid sequence.

Figure 2 shows that Northern blot analysis for TADG-12 revealed three transcripts of 2.4, 1.6 and 0.7 kilobases. These transcripts were found at significant levels in ovarian tumors and cancer cell lines, but the transcripts were found only at low levels in normal ovary.

Figure 3 shows an RNA dot blot (CLONTECH) probed for TADG-12. The transcript was detectable (at background levels) in all 50 of the human tissues represented with the greatest abundance of transcript in the heart. Putamen, amygdala, kidney, liver, small intestine, skeletal muscle, and adrenal gland were also found to have intermediate levels of TADG-12 transcript.

Figure 4 shows the entire cDNA sequence for TADG-12 (SEQ ID No. 1) with its predicted open reading frame of 454 amino acids (SEQ ID No. 2). Within the nucleotide sequence, the Kozak's consensus sequence for the initiation of translation and the poly-adenylation signal are underlined. In the protein sequence, a potential transmembrane domain is boxed. The LDLR-A domain is underlined with a solid line. The SRCR domain is underlined with a broken line. The residues of the catalytic triad of the serine protease domain are circled, and the beginning of the

catalytic domain is marked with an arrow designated as a potential proteolytic cleavage site. The * represents the stop codon that terminates translation.

Figure 5A shows the 35 amino acid LDLR-A domain of TADG-12 (SEQ ID No. 13) aligned with other LDLR-A motifs from the serine protease TMPRSS2 (U75329, SEQ ID No. 14), the complement subunit C8 (P07358, SEQ ID No. 9), two LDLR-A domains of the glycoprotein GP300 (P98164, SEQ ID Nos. 11-12), and the serine protease matriptase (AF118224, SEQ ID No. 10). TADG-12 has its highest similarity with the other serine proteases for which it is 54% similar to TMPRSS2 and 53% similar to matriptase. The highly conserved cysteine residues are shown in bold type. **Figure 5B** shows the SRCR domain of TADG-12 (SEQ ID No. 17) aligned with other domain family members including the human macrophage scavenger receptor (P21757, SEQ ID No. 16), human enterokinase (P98073, SEQ ID No. 19), bovine enterokinase (P21758, SEQ ID No. 15), and the serine protease TMPRSS2 (SEQ ID No. 18). Again, TADG-12 shows its highest similarity within this region to the protease TMPRSS2 at 43%. **Figure 5C** shows the protease domain of TADG-12 (SEQ ID No. 23) in alignment with other human serine proteases including protease M (U62801, SEQ ID No. 20), trypsinogen I (P07477, SEQ ID No. 21), plasma kallikrein (P03952, SEQ ID No. 22), hepsin (P05981, SEQ ID No. 25), and TMPRSS2 (SEQ ID No. 24). Cons represents the consensus sequence for each alignment.

Figure 6 shows semi-quantitative PCR analysis that was performed for TADG-12 (upper panel) and TADG-12V (lower panel). The amplification of TADG-12 or TADG-12V was performed in parallel with PCR amplification of β -tubulin product

as an internal control. The TADG-12 transcript was found to be overexpressed in 41 of 55 carcinomas. The TADG-12V transcript was found to be overexpressed in 8 of 22 carcinomas examined. Note that the samples in the upper panel are not necessarily the same as the samples in the lower panel.

Figure 7 shows immunohistochemical staining of normal ovary and ovarian tumors which were performed using a polyclonal rabbit antibody developed to a TADG-12 specific peptide. No significant staining was detected in normal ovary (**Figure 7A**). Strong positive staining was observed in 22 of 29 carcinomas examined. **Figures 7B and 7C** represent a serous and mucinous carcinoma, respectively. Both show diffuse staining throughout the cytoplasm of tumor cells while stromal cells remain relatively unstained.

Figure 8 is a model to demonstrate the progression of TADG-12 within a cellular context. In normal circumstances, the TADG-12 transcript is appropriately spliced and the resulting protein is capable of being expressed at the cell surface where the protease may be cleaved to an active form. The role of the remaining ligand binding domains has not yet been determined, but one can envision their potential to bind other molecules for activation, internalization or both. The TADG-12V transcript, which occurs in some tumors, may be the result of mutation and/or poor mRNA processing may be capable of producing a truncated form of TADG-12 that does not have a functional protease domain. In addition, this truncated product may present a novel epitope at the surface of tumor cells.

DETAILED DESCRIPTION OF THE INVENTION

To examine the serine proteases expressed by ovarian cancers, a PCR based differential display technique was employed
5 utilizing redundant PCR primers designed to the most highly conserved amino acids in these proteins [9]. As a result, a novel cell-surface, multi-domain serine protease, named Tumor Associated Differentially-expressed Gene-12 (TADG-12) was identified. TADG-12 appears to be overexpressed in many ovarian
10 tumors. The extracellular nature of TADG-12 may render tumors susceptible to detection via a TADG-12 specific assay. In addition, a splicing variant of TADG-12, named TADG-12V, was detected at elevated levels in 35% of the tumors that were examined. TADG-12V encodes a truncated form of TADG-12 with an altered amino
15 acid sequence that may be a unique tumor specific target for future therapeutic approaches.

The TADG-12 cDNA is 2413 base pairs long (SEQ ID No. 1) encoding a 454 amino acid protein (SEQ ID No. 2). A variant form, TADG-12V (SEQ ID No. 3), encodes a 294 amino acid protein
20 (SEQ ID No. 4). The availability of the TADG-12 and/or TADG-12V gene opens the way for a number studies that can lead to various applications. For example, the TADG-12 and/or TADG-12V gene can be used as a diagnostic or therapeutic target in ovarian carcinoma and other carcinomas including breast, prostate, lung
25 and colon.

In accordance with the present invention there may be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, e.g., Maniatis,

Fritsch & Sambrook, "Molecular Cloning: A Laboratory Manual (1982); "DNA Cloning: A Practical Approach," Volumes I and II (D.N. Glover ed. 1985); "Oligonucleotide Synthesis" (M.J. Gait ed. 1984); "Nucleic Acid Hybridization" [B.D. Hames & S.J. Higgins eds. 5 (1985)]; "Transcription and Translation" [B.D. Hames & S.J. Higgins eds. (1984)]; "Animal Cell Culture" [R.I. Freshney, ed. (1986)]; "Immobilized Cells And Enzymes" [IRL Press, (1986)]; B. Perbal, "A Practical Guide To Molecular Cloning" (1984).

Therefore, if appearing herein, the following terms
10 shall have the definitions set out below.

As used herein, the term "cDNA" shall refer to the DNA copy of the mRNA transcript of a gene.

As used herein, the term "derived amino acid sequence" shall mean the amino acid sequence determined by
15 reading the triplet sequence of nucleotide bases in the cDNA.

As used herein the term "screening a library" shall refer to the process of using a labeled probe to check whether, under the appropriate conditions, there is a sequence complementary to the probe present in a particular DNA library.
20 In addition, "screening a library" could be performed by PCR.

As used herein, the term "PCR" refers to the polymerase chain reaction that is the subject of U.S. Patent Nos. 4,683,195 and 4,683,202 to Mullis, as well as other improvements now known in the art.

25 The amino acid described herein are preferred to be in the "L" isomeric form. However, residues in the "D" isomeric form can be substituted for any L-amino acid residue, as long as the desired functional property of immunoglobulin-binding is retained by the polypeptide. NH₂ refers to the free amino group present at

the amino terminus of a polypeptide. COOH refers to the free carboxy group present at the carboxy terminus of a polypeptide. In keeping with standard polypeptide nomenclature, *J Biol. Chem.*, 243:3552-59 (1969), abbreviations for amino acid residues are
5 known in the art.

It should be noted that all amino-acid residue sequences are represented herein by formulae whose left and right orientation is in the conventional direction of amino-terminus to carboxy-terminus. Furthermore, it should be noted
10 that a dash at the beginning or end of an amino acid residue sequence indicates a peptide bond to a further sequence of one or more amino-acid residues.

A "replicon" is any genetic element (e.g., plasmid, chromosome, virus) that functions as an autonomous unit of DNA
15 replication *in vivo*; i.e., capable of replication under its own control.

A "vector" is a replicon, such as plasmid, phage or cosmid, to which another DNA segment may be attached so as to bring about the replication of the attached segment.

20 A "DNA molecule" refers to the polymeric form of deoxyribonucleotides (adenine, guanine, thymine, or cytosine) in its either single stranded form, or a double-stranded helix. This term refers only to the primary and secondary structure of the molecule, and does not limit it to any particular tertiary forms.
25 Thus, this term includes double-stranded DNA found, *inter alia*, in linear DNA molecules (e.g., restriction fragments), viruses, plasmids, and chromosomes. In discussing the structure herein according to the normal convention of giving only the sequence in

the 5' to 3' direction along the nontranscribed strand of DNA (i.e., the strand having a sequence homologous to the mRNA).

An "origin of replication" refers to those DNA sequences that participate in DNA synthesis.

5 A DNA "coding sequence" is a double-stranded DNA sequence which is transcribed and translated into a polypeptide *in vivo* when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5' (amino) terminus and a translation stop
10 codon at the 3' (carboxyl) terminus. A coding sequence can include, but is not limited to, prokaryotic sequences, cDNA from eukaryotic mRNA, genomic DNA sequences from eukaryotic (e.g., mammalian) DNA, and even synthetic DNA sequences. A polyadenylation signal and transcription termination sequence
15 will usually be located 3' to the coding sequence.

Transcriptional and translational control sequences are DNA regulatory sequences, such as promoters, enhancers, polyadenylation signals, terminators, and the like, that provide for the expression of a coding sequence in a host cell.

20 A "promoter sequence" is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3' direction) coding sequence. For purposes of defining the present invention, the promoter sequence is bounded at its 3' terminus by the transcription initiation site
25 and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence will be found a transcription initiation site, as well as protein binding domains (consensus sequences) responsible for

the binding of RNA polymerase. Eukaryotic promoters often, but not always, contain "TATA" boxes and "CAT" boxes. Prokaryotic promoters contain Shine-Dalgarno sequences in addition to the -10 and -35 consensus sequences.

5 An "expression control sequence" is a DNA sequence that controls and regulates the transcription and translation of another DNA sequence. A coding sequence is "under the control" of transcriptional and translational control sequences in a cell when RNA polymerase transcribes the coding sequence into
10 mRNA, which is then translated into the protein encoded by the coding sequence.

 A "signal sequence" can be included near the coding sequence. This sequence encodes a signal peptide, N-terminal to the polypeptide, that communicates to the host cell to direct the
15 polypeptide to the cell surface or secrete the polypeptide into the media, and this signal peptide is clipped off by the host cell before the protein leaves the cell. Signal sequences can be found associated with a variety of proteins native to prokaryotes and eukaryotes.

20 The term "oligonucleotide", as used herein in referring to the probe of the present invention, is defined as a molecule comprised of two or more ribonucleotides, preferably more than three. Its exact size will depend upon many factors which, in turn, depend upon the ultimate function and use of the oligonucleotide.

25 The term "primer" as used herein refers to an oligonucleotide, whether occurring naturally as in a purified restriction digest or produced synthetically, which is capable of acting as a point of initiation of synthesis when placed under conditions in which synthesis of a primer extension product, which

is complementary to a nucleic acid strand, is induced, i.e., in the presence of nucleotides and an inducing agent such as a DNA polymerase and at a suitable temperature and pH. The primer may be either single-stranded or double-stranded and must be sufficiently long to prime the synthesis of the desired extension product in the presence of the inducing agent. The exact length of the primer will depend upon many factors, including temperature, source of primer and use the method. For example, for diagnostic applications, depending on the complexity of the target sequence, the oligonucleotide primer typically contains 15-25 or more nucleotides, although it may contain fewer nucleotides.

The primers herein are selected to be "substantially" complementary to different strands of a particular target DNA sequence. This means that the primers must be sufficiently complementary to hybridize with their respective strands. Therefore, the primer sequence need not reflect the exact sequence of the template. For example, a non-complementary nucleotide fragment may be attached to the 5' end of the primer, with the remainder of the primer sequence being complementary to the strand. Alternatively, non-complementary bases or longer sequences can be interspersed into the primer, provided that the primer sequence has sufficient complementary with the sequence or hybridize therewith and thereby form the template for the synthesis of the extension product.

As used herein, the terms "restriction endonucleases" and "restriction enzymes" refer to enzymes, each of which cut double-stranded DNA at or near a specific nucleotide sequence.

A cell has been "transformed" by exogenous or heterologous DNA when such DNA has been introduced inside the

cell. The transforming DNA may or may not be integrated (covalently linked) into the genome of the cell. In prokaryotes, yeast, and mammalian cells for example, the transforming DNA may be maintained on an episomal element such as a plasmid.

5 With respect to eukaryotic cells, a stably transformed cell is one in which the transforming DNA has become integrated into a chromosome so that it is inherited by daughter cells through chromosome replication. This stability is demonstrated by the ability of the eukaryotic cell to establish cell lines or clones
10 comprised of a population of daughter cells containing the transforming DNA. A "clone" is a population of cells derived from a single cell or ancestor by mitosis. A "cell line" is a clone of a primary cell that is capable of stable growth *in vitro* for many generations.

15 Two DNA sequences are "substantially homologous" when at least about 75% (preferably at least about 80%, and most preferably at least about 90% or 95%) of the nucleotides match over the defined length of the DNA sequences. Sequences that are substantially homologous can be identified by comparing the
20 sequences using standard software available in sequence data banks, or in a Southern hybridization experiment under, for example, stringent conditions as defined for that particular system. Defining appropriate hybridization conditions is within the skill of the art. See, e.g., Maniatis et al., *supra*; DNA Cloning,
25 Vols. I & II, *supra*; Nucleic Acid Hybridization, *supra*.

A "heterologous" region of the DNA construct is an identifiable segment of DNA within a larger DNA molecule that is not found in association with the larger molecule in nature. Thus, when the heterologous region encodes a mammalian gene, the

gene will usually be flanked by DNA that does not flank the mammalian genomic DNA in the genome of the source organism. In another example, coding sequence is a construct where the coding sequence itself is not found in nature (e.g., a cDNA where
5 the genomic coding sequence contains introns, or synthetic sequences having codons different than the native gene). Allelic variations or naturally-occurring mutational events do not give rise to a heterologous region of DNA as defined herein.

The labels most commonly employed for these studies
10 are radioactive elements, enzymes, chemicals which fluoresce when exposed to ultraviolet light, and others. A number of fluorescent materials are known and can be utilized as labels. These include, for example, fluorescein, rhodamine, auramine, Texas Red, AMCA blue and Lucifer Yellow. A particular detecting
15 material is anti-rabbit antibody prepared in goats and conjugated with fluorescein through an isothiocyanate.

Proteins can also be labeled with a radioactive element or with an enzyme. The radioactive label can be detected by any of the currently available counting procedures. The preferred
20 isotope may be selected from ^3H , ^{14}C , ^{32}P , ^{35}S , ^{36}Cl , ^{51}Cr , ^{57}Co , ^{58}Co , ^{59}Fe , ^{90}Y , ^{125}I , ^{131}I , and ^{186}Re .

Enzyme labels are likewise useful, and can be detected by any of the presently utilized colorimetric, spectrophotometric, fluorospectrophotometric, amperometric or gasometric techniques.
25 The enzyme is conjugated to the selected particle by reaction with bridging molecules such as carbodiimides, diisocyanates, glutaraldehyde and the like. Many enzymes which can be used in these procedures are known and can be utilized. The preferred are peroxidase, β -glucuronidase, β -D-glucosidase, β -D-

galactosidase, urease, glucose oxidase plus peroxidase and alkaline phosphatase. U.S. Patent Nos. 3,654,090, 3,850,752, and 4,016,043 are referred to by way of example for their disclosure of alternate labeling material and methods.

5 A particular assay system developed and utilized in the art is known as a receptor assay. In a receptor assay, the material to be assayed is appropriately labeled and then certain cellular test colonies are inoculated with a quantity of both the label after which binding studies are conducted to determine the
10 extent to which the labeled material binds to the cell receptors. In this way, differences in affinity between materials can be ascertained.

 An assay useful in the art is known as a "cis/trans" assay. Briefly, this assay employs two genetic constructs, one of
15 which is typically a plasmid that continually expresses a particular receptor of interest when transfected into an appropriate cell line, and the second of which is a plasmid that expresses a reporter such as luciferase, under the control of a receptor/ligand complex. Thus, for example, if it is desired to evaluate a compound as a
20 ligand for a particular receptor, one of the plasmids would be a construct that results in expression of the receptor in the chosen cell line, while the second plasmid would possess a promoter linked to the luciferase gene in which the response element to the particular receptor is inserted. If the compound under test is an
25 agonist for the receptor, the ligand will complex with the receptor, and the resulting complex will bind the response element and initiate transcription of the luciferase gene. The resulting chemiluminescence is then measured photometrically, and dose response curves are obtained and compared to those of known

ligands. The foregoing protocol is described in detail in U.S. Patent No. 4,981,784.

As used herein, the term "host" is meant to include not only prokaryotes but also eukaryotes such as yeast, plant and animal cells. A recombinant DNA molecule or gene which encodes a human TADG-12 protein of the present invention can be used to transform a host using any of the techniques commonly known to those of ordinary skill in the art. Especially preferred is the use of a vector containing coding sequences for the gene which encodes a huma TADG-12 protein of the present invention for purposes of prokaryote transformation. Prokaryotic hosts may include *E. coli*, *S. typhimurium*, *Serratia marcescens* and *Bacillus subtilis*. Eukaryotic hosts include yeasts such as *Pichia pastoris*, mammalian cells and insect cells.

In general, expression vectors containing promoter sequences which facilitate the efficient transcription of the inserted DNA fragment are used in connection with the host. The expression vector typically contains an origin of replication, promoter(s), terminator(s), as well as specific genes which are capable of providing phenotypic selection in transformed cells. The transformed hosts can be fermented and cultured according to means known in the art to achieve optimal cell growth.

The invention includes a substantially pure DNA encoding a TADG-12 protein, a strand of which DNA will hybridize at high stringency to a probe containing a sequence of at least 15 consecutive nucleotides of the sequence shown in SEQ ID No. 1 or SEQ ID No. 3. The protein encoded by the DNA of this invention may share at least 80% sequence identity (preferably 85%, more preferably 90%, and most preferably 95%) with the amino acids

listed in SEQ ID No. 2 or SEQ ID No. 4. More preferably, the DNA includes the coding sequence of the nucleotides of Figure 4 (SEQ ID No. 1), or a degenerate variant of such a sequence.

5 The probe to which the DNA of the invention hybridizes preferably consists of a sequence of at least 20 consecutive nucleotides, more preferably 40 nucleotides, even more preferably 50 nucleotides, and most preferably 100 nucleotides or more (up to 100%) of the coding sequence of the nucleotides listed in Figure 4 (SEQ ID No. 1) or the complement
10 thereof. Such a probe is useful for detecting expression of TADG-12 in a human cell by a method including the steps of (a) contacting mRNA obtained from the cell with the labeled hybridization probe; and (b) detecting hybridization of the probe with the mRNA.

15 This invention also includes a substantially pure DNA containing a sequence of at least 15 consecutive nucleotides (preferably 20, more preferably 30, even more preferably 50, and most preferably all) of the region from nucleotides 1 to 2413 of the nucleotides listed in SEQ ID No. 1, or of the region from
20 nucleotides 1 to 2544 of the nucleotides listed in SEQ ID No. 3. The present invention also comprises antisense oligonucleotides directed against this novel DNA. Given the teachings of the present invention, a person having ordinary skill in this art would readily be able to develop antisense oligonucleotides directed
25 against this DNA.

By "high stringency" is meant DNA hybridization and wash conditions characterized by high temperature and low salt concentration, e.g., wash conditions of 65°C at a salt concentration of approximately 0.1 x SSC, or the functional equivalent thereof.

For example, high stringency conditions may include hybridization at about 42°C in the presence of about 50% formamide; a first wash at about 65°C with about 2 x SSC containing 1% SDS; followed by a second wash at about 65°C with about 0.1 x SSC.

5 By "substantially pure DNA" is meant DNA that is not part of a milieu in which the DNA naturally occurs, by virtue of separation (partial or total purification) of some or all of the molecules of that milieu, or by virtue of alteration of sequences that flank the claimed DNA. The term therefore includes, for
10 example, a recombinant DNA which is incorporated into a vector, into an autonomously replicating plasmid or virus, or into the genomic DNA of a prokaryote or eukaryote; or which exists as a separate molecule (e.g., a cDNA or a genomic or cDNA fragment produced by polymerase chain reaction (PCR) or restriction
15 endonuclease digestion) independent of other sequences. It also includes a recombinant DNA which is part of a hybrid gene encoding additional polypeptide sequence, e.g., a fusion protein. Also included is a recombinant DNA which includes a portion of the nucleotides shown in SEQ ID No. 3 which encodes an
20 alternative splice variant of TADG-12 (TADG-12V).

The DNA may have at least about 70% sequence identity to the coding sequence of the nucleotides listed in SEQ ID No. 1 or SEQ ID No. 3, preferably at least 75% (e.g. at least 80%); and most preferably at least 90%. The identity between two
25 sequences is a direct function of the number of matching or identical positions. When a subunit position in both of the two sequences is occupied by the same monomeric subunit, e.g., if a given position is occupied by an adenine in each of two DNA molecules, then they are identical at that position. For example, if

7 positions in a sequence 10 nucleotides in length are identical to the corresponding positions in a second 10-nucleotide sequence, then the two sequences have 70% sequence identity. The length of comparison sequences will generally be at least 50 nucleotides, preferably at least 60 nucleotides, more preferably at least 75 nucleotides, and most preferably 100 nucleotides. Sequence identity is typically measured using sequence analysis software (e.g., Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 10 1710 University Avenue, Madison, WI 53705).

The present invention comprises a vector comprising a DNA sequence which encodes a human TADG-12 protein and the vector is capable of replication in a host which comprises, in operable linkage: a) an origin of replication; b) a promoter; and c) 15 a DNA sequence coding for said protein. Preferably, the vector of the present invention contains a portion of the DNA sequence shown in SEQ ID No. 1 or SEQ ID No. 3. A "vector" may be defined as a replicable nucleic acid construct, e.g., a plasmid or viral nucleic acid. Vectors may be used to amplify and/or express 20 nucleic acid encoding a TADG-12 protein. An expression vector is a replicable construct in which a nucleic acid sequence encoding a polypeptide is operably linked to suitable control sequences capable of effecting expression of the polypeptide in a cell. The need for such control sequences will vary depending upon the cell 25 selected and the transformation method chosen. Generally, control sequences include a transcriptional promoter and/or enhancer, suitable mRNA ribosomal binding sites, and sequences which control the termination of transcription and translation. Methods which are well known to those skilled in the art can be used to

construct expression vectors containing appropriate transcriptional and translational control signals. See for example, the techniques described in Sambrook et al., 1989, *Molecular Cloning: A Laboratory Manual* (2nd Ed.), Cold Spring Harbor Press, N.Y. A gene and its transcription control sequences are defined as being "operably linked" if the transcription control sequences effectively control the transcription of the gene. Vectors of the invention include, but are not limited to, plasmid vectors and viral vectors. Preferred viral vectors of the invention are those derived from retroviruses, adenovirus, adeno-associated virus, SV40 virus, or herpes viruses.

By a "substantially pure protein" is meant a protein which has been separated from at least some of those components which naturally accompany it. Typically, the protein is substantially pure when it is at least 60%, by weight, free from the proteins and other naturally-occurring organic molecules with which it is naturally associated *in vivo*. Preferably, the purity of the preparation is at least 75%, more preferably at least 90%, and most preferably at least 99%, by weight. A substantially pure TADG-12 protein may be obtained, for example, by extraction from a natural source; by expression of a recombinant nucleic acid encoding an TADG-12 polypeptide; or by chemically synthesizing the protein. Purity can be measured by any appropriate method, e.g., column chromatography such as immunoaffinity chromatography using an antibody specific for TADG-12, polyacrylamide gel electrophoresis, or HPLC analysis. A protein is substantially free of naturally associated components when it is separated from at least some of those contaminants which accompany it in its natural state. Thus, a protein which is

chemically synthesized or produced in a cellular system different from the cell from which it naturally originates will be, by definition, substantially free from its naturally associated components. Accordingly, substantially pure proteins include
5 eukaryotic proteins synthesized in *E. coli*, other prokaryotes, or any other organism in which they do not naturally occur.

In addition to substantially full-length proteins, the invention also includes fragments (e.g., antigenic fragments) of the TADG-12 protein. As used herein, "fragment," as applied to a
10 polypeptide, will ordinarily be at least 10 residues, more typically at least 20 residues, and preferably at least 30 (e.g., 50) residues in length, but less than the entire, intact sequence. Fragments of the TADG-12 protein can be generated by methods known to those skilled in the art, e.g., by enzymatic digestion of naturally
15 occurring or recombinant TADG-12 protein, by recombinant DNA techniques using an expression vector that encodes a defined fragment of TADG-12, or by chemical synthesis. The ability of a candidate fragment to exhibit a characteristic of TADG-12 (e.g., binding to an antibody specific for TADG-12) can be assessed by
20 methods described herein. Purified TADG-12 or antigenic fragments of TADG-12 can be used to generate new antibodies or to test existing antibodies (e.g., as positive controls in a diagnostic assay) by employing standard protocols known to those skilled in the art. Included in this invention are polyclonal antisera
25 generated by using TADG-12 or a fragment of TADG-12 as the immunogen in, e.g., rabbits. Standard protocols for monoclonal and polyclonal antibody production known to those skilled in this art are employed. The monoclonal antibodies generated by this procedure can be screened for the ability to identify recombinant

TADG-12 cDNA clones, and to distinguish them from known cDNA clones.

Further included in this invention are TADG-12 proteins which are encoded at least in part by portions of SEQ ID No. 1 or SEQ ID No. 3, e.g., products of alternative mRNA splicing or alternative protein processing events, or in which a section of TADG-12 sequence has been deleted. The fragment, or the intact TADG-12 polypeptide, may be covalently linked to another polypeptide, e.g. which acts as a label, a ligand or a means to increase antigenicity.

The invention also includes a polyclonal or monoclonal antibody which specifically binds to TADG-12. The invention encompasses not only an intact monoclonal antibody, but also an immunologically-active antibody fragment, e.g., a Fab or (Fab)₂ fragment; an engineered single chain Fv molecule; or a chimeric molecule, e.g., an antibody which contains the binding specificity of one antibody, e.g., of murine origin, and the remaining portions of another antibody, e.g., of human origin.

In one embodiment, the antibody, or a fragment thereof, may be linked to a toxin or to a detectable label, e.g. a radioactive label, non-radioactive isotopic label, fluorescent label, chemiluminescent label, paramagnetic label, enzyme label, or colorimetric label. Examples of suitable toxins include diphtheria toxin, *Pseudomonas* exotoxin A, ricin, and cholera toxin. Examples of suitable enzyme labels include malate hydrogenase, staphylococcal nuclease, delta-5-steroid isomerase, alcohol dehydrogenase, alpha-glycerol phosphate dehydrogenase, triose phosphate isomerase, peroxidase, alkaline phosphatase, asparaginase, glucose oxidase, beta-galactosidase, ribonuclease,

urease, catalase, glucose-6-phosphate dehydrogenase, glucoamylase, acetylcholinesterase, etc. Examples of suitable radioisotopic labels include ^3H , ^{125}I , ^{131}I , ^{32}P , ^{35}S , ^{14}C , etc.

Paramagnetic isotopes for purposes of *in vivo* diagnosis can also be used according to the methods of this invention. There are numerous examples of elements that are useful in magnetic resonance imaging. For discussions on *in vivo* nuclear magnetic resonance imaging, see, for example, Schaefer et al., (1989) *JACC* 14, 472-480; Shreve et al., (1986) *Magn. Reson. Med.* 3, 336-340; Wolf, G. L., (1984) *Physiol. Chem. Phys. Med. NMR* 16, 93-95; Wesbey et al., (1984) *Physiol. Chem. Phys. Med. NMR* 16, 145-155; Runge et al., (1984) *Invest. Radiol.* 19, 408-415. Examples of suitable fluorescent labels include a fluorescein label, an isothiocyalate label, a rhodamine label, a phycoerythrin label, a phycocyanin label, an allophycocyanin label, an ophthaldehyde label, a fluorescamine label, etc. Examples of chemiluminescent labels include a luminal label, an isoluminal label, an aromatic acridinium ester label, an imidazole label, an acridinium salt label, an oxalate ester label, a luciferin label, a luciferase label, an aequorin label, etc.

Those of ordinary skill in the art will know of other suitable labels which may be employed in accordance with the present invention. The binding of these labels to antibodies or fragments thereof can be accomplished using standard techniques commonly known to those of ordinary skill in the art. Typical techniques are described by Kennedy et al., (1976) *Clin. Chim. Acta* 70, 1-31; and Schurs et al., (1977) *Clin. Chim. Acta* 81, 1-40. Coupling techniques mentioned in the latter are the glutaraldehyde method, the periodate method, the dimaleimide

method, the m-maleimidobenzyl-N-hydroxy-succinimide ester method. All of these methods are incorporated by reference herein.

Also within the invention is a method of detecting
5 TADG-12 protein in a biological sample, which includes the steps of contacting the sample with the labeled antibody, e.g., radioactively tagged antibody specific for TADG-12, and determining whether the antibody binds to a component of the sample.

10 As described herein, the invention provides a number of diagnostic advantages and uses. For example, the TADG-12 protein disclosed in the present invention is useful in diagnosing cancer in different tissues since this protein is highly overexpressed in tumor cells. Antibodies (or antigen-binding
15 fragments thereof) which bind to an epitope specific for TADG-12, are useful in a method of detecting TADG-12 protein in a biological sample for diagnosis of cancerous or neoplastic transformation. This method includes the steps of obtaining a biological sample (e.g., cells, blood, plasma, tissue, etc.) from a patient suspected of
20 having cancer, contacting the sample with a labeled antibody (e.g., radioactively tagged antibody) specific for TADG-12, and detecting the TADG-12 protein using standard immunoassay techniques such as an ELISA. Antibody binding to the biological sample indicates that the sample contains a component which specifically
25 binds to an epitope within TADG-12.

Likewise, a standard Northern blot assay can be used to ascertain the relative amounts of TADG-12 mRNA in a cell or tissue obtained from a patient suspected of having cancer, in accordance with conventional Northern hybridization techniques

known to those of ordinary skill in the art. This Northern assay uses a hybridization probe, e.g. radiolabelled TADG-12 cDNA, either containing the full-length, single stranded DNA having a sequence complementary to SEQ ID No. 1 or SEQ ID No. 3, or a
5 fragment of that DNA sequence at least 20 (preferably at least 30, more preferably at least 50, and most preferably at least 100 consecutive nucleotides in length). The DNA hybridization probe can be labeled by any of the many different methods known to those skilled in this art.

10 Antibodies to the TADG-12 protein can be used in an immunoassay to detect increased levels of TADG-12 protein expression in tissues suspected of neoplastic transformation. These same uses can be achieved with Northern blot assays and analyses.

15 The present invention is directed to DNA fragment encoding a TADG-12 protein selected from the group consisting of:
(a) an isolated DNA fragment which encodes a TADG-12 protein;
(b) an isolated DNA fragment which hybridizes to isolated DNA fragment of (a) above and which encodes a TADG-12 protein; and
20 (c) an isolated DNA fragment differing from the isolated DNA fragments of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-12 protein. Preferably, the DNA has the sequence shown in SEQ ID No. 1 or SEQ ID No. 3. More preferably, the DNA encodes a TADG-
25 12 protein having the amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4.

The present invention is also directed to a vector and/or a host cell capable of expressing the DNA of the present invention. Preferably, the vector contains DNA encoding a TADG-

12 protein having the amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4. Representative host cells include bacterial cells, yeast cells, mammalian cells and insect cells.

5 The present invention is also directed to an isolated and purified TADG-12 protein coded for by DNA selected from the group consisting of: (a) isolated DNA which encodes a TADG-12 protein; (b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-12 protein; and (c) isolated DNA
10 differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-12 protein. Preferably, the isolated and purified TADG-12 protein has the amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4.

15 The present invention is also directed to a method of detecting expression of the TADG-12 protein described herein, comprising the steps of: (a) contacting mRNA obtained from the cell with the labeled hybridization probe; and (b) detecting hybridization of the probe with the mRNA.

20 A number of potential applications are possible for the TADG-12 gene and gene product including the truncated product TADG-12V.

25 In one embodiment of the present invention, there is provided a method for diagnosing a cancer by detecting a TADG-12 protein in a biological sample, wherein the presence or absence of a TADG-12 protein indicates the presence or absence of a cancer. Preferably, the biological sample is selected from the group consisting of blood, urine, saliva, tears, interstitial fluid, ascites fluid, tumor tissue biopsy and circulating tumor cells. Still preferably, the detection of TADG-12 protein is by means selected

from the group consisting of Northern blot, Western blot, PCR, dot blot, ELIZA sandwich assay, radioimmunoassay, DNA array chips and flow cytometry. Such method is used for detecting an ovarian cancer, breast cancer, lung cancer, colon cancer, prostate cancer
5 and other cancers in which TADG-12 is overexpressed.

In another embodiment of the present invention, there is provided a method for detecting malignant hyperplasia by detecting a TADG-12 protein or TADG-12 mRNA in a biological sample. Further by comprising the TADG-12 protein or TADG-12
10 mRNA to reference information, a diagnosis or a treatment can be provided. Preferably, PCR amplification is used for detecting TADG-12 mRNA, wherein the primers utilized are selected from the group consisting of SEQ ID Nos. 28-31. Still preferably, detection of a TADG-12 protein is by immunoaffinity to an
15 antibody directed against a TADG-12 protein.

In still another embodiment of the present invention, there is provided a method of inhibiting expression of endogenous TADG-12 mRNA in a cell by introducing a vector comprising a DNA fragment of TADG-12 in opposite orientation operably linked to
20 elements necessary for expression. As a result, the vector produces TADG-12 antisense mRNA in the cell, which hybridizes to endogenous TADG-12 mRNA, thereby inhibiting expression of endogenous TADG-12 mRNA.

In still yet another embodiment of the present
25 invention, there is provided a method of inhibiting expression of a TADG-12 protein by introducing an antibody directed against a TADG-12 protein or fragment thereof. As a result, the binding of the antibody to the TADG-12 protein or fragment thereof inhibits the expression of the TADG-12 protein.

TADG-12 gene products including the truncated form can be used for targeted therapy. Specifically, a compound having a targeting moiety specific for a TADG-12 protein and a therapeutic moiety is administered to an individual in need of such treatment. Preferably, the targeting moiety is selected from the group consisting of an antibody directed against a TADG-12 protein and a ligand or ligand binding domain that binds a TADG-12 protein. The TADG-12 protein has an amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4. Still preferably, the therapeutic moiety is selected from the group consisting of a radioisotope, a toxin, a chemotherapeutic agent, an immune stimulant and a cytotoxic agent. Such method can be used for treating an individual having a disease selected from the group consisting of ovarian cancer, lung cancer, prostate cancer, colon cancer and other cancers in which TADG-12 is overexpressed.

In yet another embodiment of the present invention, there is provided a method of vaccinating, or producing an immune response in, an individual against TADG-12 by inoculating the individual with a TADG-12 protein or fragment thereof. Specifically, the TADG-12 protein or fragment thereof lacks TADG-12 activity, and the inoculation elicits an immune response in the individual, thereby vaccinating the individual against TADG-12. Preferably, the individual has a cancer, is suspected of having a cancer or is at risk of getting a cancer. Still preferably, TADG-12 protein has an amino acid sequence shown in SEQ ID No. 2 or SEQ ID No. 4, while TADG-12 fragment has a sequence shown in SEQ ID No. 8, or is a 9-residue fragment up to a 20-residue fragment. Examples of 9-residue fragment are shown in SEQ ID Nos. 35, 36, 55, 56, 83, 84, 97, 98, 119, 120, 122, 123 and 136.

In still yet another embodiment of the present invention, there is provided an immunogenic composition, comprising an immunogenic fragment of a TADG-12 protein and an appropriate adjuvant. Preferably, the immunogenic fragment
5 of the TADG-12 protein has a sequence shown in SEQ ID No. 8, or is a 9-residue fragment up to a 20-residue fragment. Examples of 9-residue fragment are shown in SEQ ID Nos. 35, 36, 55, 56, 83, 84, 97, 98, 119, 120, 122, 123 and 136.

The following examples are given for the purpose of
10 illustrating various embodiments of the invention and are not meant to limit the present invention in any fashion.

EXAMPLE 1

Tissue collection and storage

15 Upon patient hysterectomy, bilateral salpingo-oophorectomy, or surgical removal of neoplastic tissue, the specimen is retrieved and placed on ice. The specimen was then taken to the resident pathologist for isolation and identification of specific tissue samples. Finally, the sample was frozen in liquid
20 nitrogen, logged into the laboratory record and stored at -80°C. Additional specimens were frequently obtained from the Cooperative Human Tissue Network (CHTN). These samples were prepared by the CHTN and shipped on dry ice. Upon arrival, these specimens were logged into the laboratory record and stored at -
25 80°C.

EXAMPLE 2

mRNA Extraction and cDNA Synthesis

Sixty-nine ovarian tumors (4 benign tumors, 10 low malignant potential tumors and 55 carcinomas) and 10 normal

ovaries were obtained from surgical specimens and frozen in liquid nitrogen. The human ovarian carcinoma cell lines SW 626 and Caov 3, the human breast carcinoma cell lines MDA-MB-231 and MDA-MB-435S were purchased from the American Type Culture Collection (Rockville, MD). Cells were cultured to sub-confluency in Dulbecco's modified Eagle's medium, supplemented with 10% (v/v) fetal bovine serum and antibiotics.

Extraction of mRNA and cDNA synthesis were carried out by the methods described previously [14-16]. mRNA was isolated by using a RiboSep mRNA isolation kit (Becton Dickinson Labware). In this procedure, poly A⁺ mRNA was isolated directly from the tissue lysate using the affinity chromatography media oligo(dT) cellulose. cDNA was synthesized with 5.0 µg of mRNA by random hexamer priming using 1st strand cDNA synthesis kit (CLONTECH).

EXAMPLE 3

PCR with Redundant Primers and Cloning of TADG-12 cDNA

Redundant primers, forward 5'-
 20 TGGGTIGTIACIGCIGCICA(CT)TG -3' (SEQ ID No. 26) and reverse 5'-
 A(AG)IA(AG)IGCIATITCITTICC-3' (SEQ ID No. 27), for the
 consensus sequences of amino acids surrounding the catalytic
 triad for serine proteases were used to compare the PCR products
 from normal and carcinoma cDNAs. The appropriate bands were
 25 ligated into Promega T-vector plasmid and the ligation product
 was used to transform JM109 cells (Promega) grown on selection
 media. After selection of individual colonies, they were cultured
 and plasmid DNA was isolated by means of the Wizard miniprep
 DNA purification system (Promega). Nucleotide sequencing was

performed using PRISM Ready Reaction Dye Deoxy terminator cycle sequencing kit (Applied Biosystems). Applied Biosystems Model 373A DNA sequencing system was used for direct cDNA sequence determination.

5 The original TADG-12 subclone was randomly labeled and used as a probe to screen an ovarian tumor cDNA library by standard hybridization techniques [11,15]. The library was constructed in λ ZAP using mRNA isolated from the tumor cells of a stage III/grade III ovarian adenocarcinoma patient. Three
10 overlapping clones were obtained which spanned 2315 nucleotides. The final 99 nucleotides encoding the most 3' sequence including the poly A tail was identified by homology with clones available in the GenBank EST database.

15

EXAMPLE 4

Quantitative PCR

The mRNA overexpression of TADG-12 was determined using a quantitative PCR. Quantitative PCR was performed according to the procedure as previously reported [16].
20 Oligonucleotide primers were used for: TADG-12, forward 5'-GAAACATGTCCTTGCTCTCG-3' (SEQ ID No. 28) and reverse 5'-ACTAACTTCCACAGCCTCCT-3' (SEQ ID No. 29); the variant TADG-12, forward 5'-TCCAGGTGGGTCTAGTTTCC-3' (SEQ ID No. 30), reverse 5'-CTCTTTGGCTTGTA CT TGCT-3' (SEQ ID No. 31); β -tubulin, forward
25 5'-CGCATCAACGTGTACTACAA-3' (SEQ ID No. 32) and reverse 5'-TACGAGCTGGTGGACTGAGA-3' (SEQ ID No. 33). β -tubulin was utilized as an internal control. The PCR reaction mixture consists of cDNA derived from 50 ng of mRNA, 5 pmol of sense and antisense primers for both the TADG-12 gene and the β -tubulin

gene, 200 μ mol of dNTPs, 5 μ Ci of α -³²PdCTP and 0.25 unit of Taq DNA polymerase with reaction buffer (Promega) in a final volume of 25 μ l. The target sequences were amplified in parallel with the β -tubulin gene. Thirty cycles of PCR were carried out in a Thermal
5 Cycler (Perkin-Elmer Cetus). Each cycle of PCR included 30 seconds of denaturation at 94°C, 30 seconds of annealing at 60°C and 30 seconds of extension at 72°C. The PCR products were separated on 2% agarose gels and the radioactivity of each PCR product was determined by using a Phospho Imager (Molecular
10 Dynamics). The present study used the expression ratio (TADG-12/ β -tubulin) as measured by phosphoimager to evaluate gene expression and defined the value at mean + 2SD of normal ovary as the cut-off value to determine overexpression. The student's *t* test was used for comparison of the mean values of normal ovary
15 and tumors.

EXAMPLE 5

Sequencing of TADG-12/TADG-12V

Utilizing a plasmid specific primer near the cloning
20 site, sequencing reactions were carried out using PRISM™ Ready Reaction Dye Deoxy™ terminators (Applied Biosystems cat# 401384) according to the manufacturer's instructions. Residual dye terminators were removed from the completed sequencing reaction using a Centri-sep™ spin column (Princeton Separation
25 cat.# CS-901). An Applied Biosystems Model 373A DNA Sequencing System was available and was used for sequence analysis.

EXAMPLE 6

Antibody Production

Polyclonal rabbit antibodies were generated by immunization of white New Zealand rabbits with a poly-lysine
5 linked multiple antigen peptide derived from the TADG-12
carboxy-terminal protein sequence NH_2 -WIHEQMERDLKT-COOH
(WIHEQMERDLKT, SEQ ID No. 34). This peptide is present in full
length TADG-12, but not TADG-12V. Rabbits were immunized
with approximately 100 μg of peptide emulsified in Ribi adjuvant.
10 Subsequent boost immunizations were carried out at 3 and 6
weeks, and rabbit serum was isolated 10 days after the boost
inoculations. Sera were tested by dot blot analysis to determine
affinity for the TADG-12 specific peptide. Rabbit pre-immune
serum was used as a negative control.

15

EXAMPLE 7

Northern Blot Analysis

10 μg of mRNA were loaded onto a 1% formaldehyde-
agarose gel, electrophoresed and blotted on a Hybond-N+ nylon
20 membrane (Amersham). ^{32}P -labeled cDNA probes were made by
Prime-a-Gene Labeling System (Promega). The PCR products
amplified by the same primers as above were used for probes.
The blots were prehybridized for 30 min and hybridized for 60
min at 68°C with ^{32}P -labeled cDNA probe in ExpressHyb
25 Hybridization Solution (CLONTECH). Control hybridization to
determine relative gel loading was performed with the β -tubulin
probe.

Normal human tissues; spleen, thymus, prostate, testis, ovary, small intestine, colon and peripheral blood leukocyte, and normal human fetal tissues; brain, lung, liver and kidney (Human Multiple Tissue Northern Blot; CLONTECH) were also examined by
5 same hybridization procedure.

EXAMPLE 8

Immunohistochemistry

Immunohistochemical staining was performed using a
10 Vectastain Elite ABC Kit (Vector). Formalin fixed and paraffin embedded specimens were routinely deparaffinized and processed using microwave heat treatment in 0.01 M sodium citrate buffer (pH 6.0). The specimens were incubated with normal goat serum in a moist chamber for 30 minutes. TADG-12 peptide antibody
15 was allowed to incubate with the specimens in a moisture chamber for 1 hour. Excess antibody was washed away with phosphate buffered saline. After incubation with biotinylated anti-rabbit IgG for 30 minutes, the sections were then incubated with ABC reagent (Vector) for 30 minutes. The final products
20 were visualized using the AEC substrate system (DAKO) and sections were counterstained with hematoxylin before mounting. Negative controls were performed by using normal serum instead of the primary antibody.

25

EXAMPLE 9

Isolation of Catalytic Domain Subclones of TADG-12 and TADG-12 Variant

To identify serine proteases that are expressed in ovarian tumors, redundant PCR primers designed to the conserved

regions of the catalytic triad of these enzymes were employed. A sense primer designed to the region surrounding the conserved histidine and an anti-sense primer designed to the region surrounding the conserved aspartate were used in PCR reactions with either normal ovary or ovarian tumor cDNA as template. In the reaction with ovarian tumor cDNA, a strong product band of the expected size of approximately 180 bp was observed as well as an unexpected PCR product of approximately 300 bp which showed strong expression in some ovarian tumor cDNA's (Figure 1A). Both of these PCR products were subcloned and sequenced. The sequence of the subclones from the 180bp band (SEQ ID No. 5) was found to be homologous to the sequence identified in the larger, unexpected band (SEQ ID No. 7) except that the larger band had an additional insert of 133 nucleotides (Figure 1B). The smaller product of the appropriate size encoded for a protein sequence (SEQ ID No. 6) homologous to other known proteases while the sequence with the insertion (SEQ ID No. 8) encoded for a frame shift from the serine protease catalytic domain and a subsequent premature translational stop codon. TADG-12 variants from four individual tumors were also subcloned and sequenced. It was found that the sequence and insert to be identical. The genomic sequences for these cDNA derived clones were amplified by PCR, examined and found to contain potential AG/GT splice sites that would allow for the variant transcript production.

25

EXAMPLE 10

Northern Blot Analysis of TADG-12 Expression

To examine transcript size and tissue distribution, the catalytic domain subclone was randomly labeled and used to

probe Northern blots representing normal ovarian tissue, ovarian tumors and the cancer cell lines SW626, CAOV3, HeLa, MD-MBA-435S and MD-MBA-231 (Figure 2). Three transcripts of 2.4, 1.6 and 0.7 kilobases were observed. In blots of normal and ovary tumor the smallest transcript size 0.7 kb was lowly expressed in normal ovary while all transcripts (2.4, 1.6 and 0.7 kb) were abundantly present in serous carcinoma. In addition, Northern blots representing the normal human tissues spleen, thymus, prostate, testis, ovary, small intestine, colon and peripheral blood leukocyte, and normal human fetal tissues of brain, lung, liver and kidney were examined. The same three transcripts were found to be expressed weakly in all of these tissues (data not shown). A human β -tubulin specific probe was utilized as a control for relative sample loading. In addition, an RNA dot blot was probed representing 50 human tissues and determined that this clone is weakly expressed in all tissues represented (Figure 3). It was found most prominently in heart, with intermediate levels in putamen, amygdala, kidney, liver, small intestine, skeletal muscle, and adrenal gland.

20

EXAMPLE 11

Sequencing and Characterization of TADG-12

An ovarian tumor cDNA library constructed in λ ZAP was screened by standard hybridization techniques using the catalytic domain subclone as a probe. Two clones that overlapped with the probe were identified and sequenced and found to represent 2316 nucleotides. The 97 nucleotides at the 3' end of the transcript including the poly-adenylation signal and the poly (A) tail were identified by homology with clones available in

GenBank's EST database. This brought the total size of the transcript to 2413 bases (SEQ ID No. 1, Figure 4). Subsequent screening of GenBank's Genomic Database revealed that TADG-12 is homologous to a cosmid from chromosome 17. This cosmid has the accession number AC015555.

The identified cDNA includes an open reading frame that would produce a predicted protein of 454 amino acids (SEQ ID No. 2), named Tumor Associated Differentially-Expressed Gene 12 (TADG-12). The sequence has been submitted to the GenBank database and granted the accession # AF201380. Using homology alignment programs, this protein contains several domains including an amino-terminal cytoplasmic domain, a potential Type II transmembrane domain followed by a low-density lipoprotein receptor-like class A domain (LDLR-A), a scavenger receptor cysteine rich domain (SRCR), and an extracellular serine protease domain.

As predicted by the TMPred program, TADG-12 contains a highly hydrophobic stretch of amino acids that could serve as a potential transmembrane domain, which would retain the amino terminus of the protein within the cytoplasm and expose the ligand binding domains and protease domain to the extracellular space. This general structure is consistent with other known transmembrane proteases including hepsin [17], and TMPRSS2 [18], and TADG-12 is particularly similar in structure to the TMPRSS2 protease.

The LDLR-A domain of TADG-12 is represented by the sequence from amino acid 74 to 108 (SEQ ID No. 13). The LDLR-A domain was originally identified within the LDL Receptor [19] as a series of repeated sequences of approximately 40 amino acids,

which contained 6 invariant cysteine residues and highly conserved aspartate and glutamate residues. Since that initial identification, a host of other genes have been identified which contain motifs homologous to this domain [20]. Several proteases
5 have been identified which contain LDLR-A motifs including matriptase, TMPRSS2 and several complement components. A comparison of TADG-12 with other known LDLR-A domains is shown in Figure 5A. The similarity of these sequences range from 44 to 54% of similar or identical amino acids.

10 In addition to the LDLR-A domain, TADG-12 contains another extracellular ligand binding domain with homology to the group A SRCR family. This family of protein domains typically is defined by the conservation of 6 cysteine residues within a sequence of approximately 100 amino acids [23]. The SRCR
15 domain of TADG-12 is encoded by amino acids 109 to 206 (SEQ ID No. 17), and this domain was aligned with other SRCR domains and found to have between 36 and 43% similarity (Figure 5B). However, TADG-12 only has 4 of the 6 conserved cysteine residues. This is similar to the SRCR domain found in the protease
20 TMPRSS2.

The TADG-12 protein also includes a serine protease domain of the trypsin family of proteases. An alignment of the catalytic domain of TADG-12 with other known proteases is shown in Figure 5C. The similarity among these sequence ranges from 48
25 to 55%, and TADG-12 is most similar to the serine protease TMPRSS2 which also contains a transmembrane domain, LDLR-A domain and an SRCR domain. There is a conserved amino acid motif (RIVGG) downstream from the SRCR domain that is a potential cleavage/activation site common to many serine

proteases of this family [25]. This suggests that TADG-12 is trafficked to the cell surface where the ligand binding domains are capable of interacting with extracellular molecules and the protease domain is potentially activated. TADG-12 also contains
5 conserved cysteine residues (amino acids 208 and 243) which in other proteases form a disulfide bond capable of linking the activated protease to the other extracellular domains.

EXAMPLE 12

10 Quantitative PCR Characterization of the Alternative Transcript

The original TADG-12 subclone was identified as highly expressed in the initial redundant-primer PCR experiment. The TADG-12 variant form (TADG-12V) with the insertion of 133 bp was also easily detected in the initial experiment. To identify
15 the frequency of this expression and whether or not the expression level between normal ovary and ovarian tumors was different, a previously authenticated semi-quantitative PCR technique was employed [16]. The PCR analysis co-amplified a product for β -tubulin with either a product specific to TADG-12 or
20 TADG-12V in the presence of a radiolabelled nucleotide. The products were separated by agarose gel electrophoresis and a phosphoimager was used to quantitate the relative abundance of each PCR product. Examples of these PCR amplification products are shown for both TADG-12 and TADG-12V in Figure 6. Normal
25 expression was defined as the mean ratio of TADG-12 (or TADG-12V) to β -tubulin \pm 2SD as examined in normal ovarian samples. For tumor samples, overexpression was defined as $>2SD$ from the normal TADG-12/ β -tubulin or TADG-12V/ β -tubulin ratio. The results are summarized in Table 1 and Table 2. TADG-12 was

found to be overexpressed in 41 of 55 carcinomas examined while the variant form was present at aberrantly high levels in 8 of 22 carcinomas. As determined by the student's t test, these differences were statistically significant ($p < 0.05$).

5

TABLE 1**Frequency of Overexpression of TADG-12 in Ovarian Carcinoma**

Histology Type	TADG-12 (%)
Normal	0/16 (0%)
LMP-Serous	3/6 (50%)
LMP-Mucinous	0/4 (0%)
Serous Carcinoma	23/29 (79%)
Mucinous Carcinoma	7/12 (58%)
Endometrioid Carcinoma	8/8 (100%)
Clear Cell Carcinoma	3/6 (50%)
Benign Tumors	3/4 (75%)

10

Overexpression = more than two standard deviations above the mean for normal ovary

LMP = low malignant potential tumor

TABLE 2**Frequency of Overexpression of TADG-12V in Ovarian Carcinoma**

Histology Type	TADG-12V (%)
Normal	0/10 (0%)
LMP-Serous	0/5 (0%)
LMP-Mucinous	0/3 (0%)
Serous Carcinoma	4/14 (29%)
Mucinous Carcinoma	3/5 (60%)
Endometrioid Carcinoma	1/3 (33%)
Clear Cell Carcinoma	N/D

Overexpression = more than two standard deviations above
 5 the mean for normal ovary; LMP = low malignant potential tumor

EXAMPLE 13**Immunohistochemical Analysis of TADG-12 in Ovarian Tumor Cells**

10 In order to examine the TADG-12 protein, polyclonal rabbit anti-sera to a peptide located in the carboxy-terminal amino acid sequence was developed. These antibodies were used to examine the expression level of the TADG-12 protein and its localization within normal ovary and ovarian tumor cells by
 15 immuno-localization. No staining was observed in normal ovarian tissues (Figure 7A) while significant staining was observed in 22 of 29 tumors studied. Representative tumor samples are shown in Figures 7B and 7C. It should be noted that TADG-12 is found in a diffuse pattern throughout the cytoplasm indicative of a protein in
 20 a trafficking pathway. TADG-12 is also found at the cell surface in these tumor samples as expected. It should be noted that the

antibody developed and used for immunohistochemical analysis would not detect the TADG-12V truncated protein.

The results of the immunohistochemical staining are summarized in Table 3. 22 of 29 ovarian tumors showed positive staining of TADG-12, whereas normal ovarian surface epithelium showed no expression of the TADG-12 antigen. 8 of 10 serous adenocarcinomas, 8 of 8 mucinous adenocarcinomas, 1 of 2 clear cell carcinomas, and 4 of 6 endometrioid carcinomas showed positive staining.

TABLE 3

Case	Stage	Histology	Grade	LN*	TADG12	Prognosis
1		Normal ovary			0 -	
2		Normal ovary			0 -	
3		Normal ovary			0 -	
4		Mucinous B		ND	0 -	Alive
5		Mucinous B		ND	1+	Alive
6	1 a	Serous LMP	G1	ND	1+	Alive
7	1 a	Mucinous LMP	G1	ND	1+	Alive
8	1 a	Mucinous CA	G1	ND	1+	Alive
9	1 a	Mucinous CA	G2	ND	1+	Alive
10	1 a	Endometrioid CA	G1	ND	0 -	Alive
11	1 c	Serous CA	G1	N	1+	Alive
12	1 c	Mucinous CA	G1	N	1+	Alive
13	1 c	Mucinous CA	G1	N	2+	Alive
14	1 c	Clear cell CA	G2	N	0 -	Alive
15	1 c	Clear cell CA	G2	N	0 -	Alive
16	2 c	Serous CA	G3	N	2+	Alive
17	3 a	Mucinous CA	G2	N	2+	Alive

18	3 b	Serous CA	G1	ND	1+	Alive
19	3c	Serous CA	G1	N	0 -	Dead
20	3c	Serous CA	G3	P	1+	Alive
21	3c	Serous CA	G2	P	2+	Alive
22	3c	Serous CA	G1	P	2+	Unknown
23	3c	Serous CA	G3	ND	2+	Alive
24	3c	Serous CA	G2	N	0 -	Dead
25	3c	Mucinous CA	G1	P	2+	Dead
26	3c	Mucinous CA	G2	ND	1+	Unknown
27	3c	Mucinous CA	G2	N	1+	Alive
28	3c	Endometrioid CA	G1	P	1+	Dead
29	3c	Endometrioid CA	G2	N	0 -	Alive
30	3c	Endometrioid CA	G2	P	1+	Dead
31	3c	Endometrioid CA	G3	P	1+	Alive
32	3c	Clear Cell CA	G3	P	2+	Dead

LN*= Lymph Node: B = Benign; N = Negative; P = Positive;

ND = Not Done

5

EXAMPLE 14

Peptide Ranking

For vaccine or immune stimulation, individual 9-mers to 11-mers of the TADG-12 protein were examined to rank the binding of individual peptides to the top 8 haplotypes in the general population [Parker et al., (1994)]. The computer program used for this analysis can be found at <http://www-bimas.dcrn.nih.gov/molbio/hla_bind/>. Table 4 shows the peptide ranking based upon the predicted half-life of each peptide's binding to a particular HLA allele. A larger half-life indicates a

stronger association with that peptide and the particular HLA molecule. The TADG-12 peptides that strongly bind to an HLA allele are putative immunogens, and are used to inoculate an individual against TADG-12.

5

TABLE 4

TADG-12 peptide ranking						
HLA Type & Ranking			Start	Peptide	Predicted Dissociation _{1/2}	SEQ ID No.
10	HLA A0201					
	1	40	ILSLLPFEV	685.783	35	
	2	144	AQLGFPSYV	545.316	36	
	3	225	LLSQWPWQA	63.342	37	
	4	252	WIITAAHCV	43.992	38	
15	5	356	VLNHAAVPL	36.316	39	
	6	176	LLPDDKVTA	34.627	40	
	7	13	FSFRSLFGL	31.661	41	
	8	151	YVSSDNLRV	27.995	42	
	9	436	RVTSFLDWI	21.502	43	
20	10	234	SLQFQGYHL	21.362	44	
	11	181	KVTALHHSV	21.300	45	
	12	183	TALHHSVYV	19.658	46	
	13	411	RLWKLVGAT	18.494	47	
	14	60	LILALAIGL	18.476	48	
25	15	227	SQWPWQASL	17.977	49	
	16	301	RLGNDIALM	11.426	50	
	17	307	ALMKLAGPL	10.275	51	
	18	262	DLYLPKSWT	9.837	52	
	19	416	LVGATSEGI	9.001	53	
30	20	54	SLGIIALIL	8.759	54	

HLA A0205

	1	218	IVGGNMSLL	47.600	55
	2	60	LILALAIGL	35.700	48
	3	35	AVAAQILSL	28.000	56
5	4	307	ALMKLAGPL	21.000	51
	5	271	IQVGLVSL	19.040	57
	6	397	CQGDSGGPL	16.800	58
	7	227	SQWPWQASL	16.800	49
	8	270	TIQVGLVSL	14.000	59
10	9	56	GIIALILAL	14.000	60
	10	110	RVGGQNAVL	14.000	61
	11	181	KVTALHHSV	12.000	45
	12	151	YVSSDNLRV	12.000	42
	13	356	VLNHAAVPL	11.900	39
15	14	144	AQLGFPSYV	9.600	36
	15	13	FSFRSLFGL	7.560	41
	16	54	SLGIIALIL	7.000	54
	17	234	SLQFQGYHL	7.000	44
	18	217	RIVGGNMSL	7.000	62
20	19	411	RLWKLVGAT	6.000	47
	20	252	WIITAAHCV	6.000	38

HLA A1

	1	130	CSDDWKGHY	37.500	63
	2	8	AVEAPFSFR	9.000	64
25	3	328	NSEENFPDG	2.700	65
	4	3	ENDPPAVEA	2.500	66
	5	98	DCKDGEDEY	2.500	67
	6	346	ATEDGGDAS	2.250	68
	7	360	AAVPLISNK	2.000	69

	8	153	SSDNLRVSS	1.500	70
	9	182	VTALHHSVY	1.250	71
	10	143	CAQLGFPSY	1.000	72
	11	259	CVYDLYLPK	1.000	73
5	12	369	ICNHRDVYG	1.000	74
	13	278	LLDNPAPSH	1.000	75
	14	426	CAEVNKPVG	1.000	76
	15	32	DADAVAAQI	1.000	77
	16	406	VCQERRLWK	1.000	78
10	17	329	SEENFPDGK	0.900	79
	18	303	GNDIALMKL	0.625	80
	19	127	KTMCSDDWK	0.500	81
	20	440	FLDWIHEQM	0.500	82
HLA A24					
15	1	433	VYTRVTSFL	280.000	83
	2	263	LYLPKSWTI	90.000	84
	3	169	EFVSIDHLL	42.000	85
	4	217	RIVGGNMSL	12.000	62
	5	296	KYKPKRLGN	12.000	86
20	6	16	RSLFGLDDL	12.000	87
	7	267	KSWTIQVGL	11.200	88
	8	81	RSSFKCIEL	8.800	89
	9	375	VYGGIISPS	8.000	90
	10	110	RVGGQNAVL	8.000	91
25	11	189	VYVREGCAS	7.500	92
	12	60	LILALAIGL	7.200	48
	13	165	QFREEFVSI	7.200	93
	14	271	IQVGLVSL	7.200	57
	15	56	GIIALILAL	7.200	60

	16	10	EAPFSFRSL	7.200	94
	17	307	ALMKLAGPL	7.200	51
	18	407	CQERRLWKL	6.600	95
	19	356	VLNHAAPVPL	6.000	39
5	20	381	SPSMLCAGY	6.000	96
HLA B7					
	1	375	VYGGIISPS	200.000	97
	2	381	SPSMLCAGY	80.000	98
	3	362	VPLISNKIC	80.000	99
10	4	35	AVAAQILSL	60.000	56
	5	373	RDVYGGIIS	40.000	100
	6	307	ALMKLAGPL	36.000	51
	7	283	APSHLVEKI	24.000	101
	8	177	LPDDKV TAL	24.000	102
15	9	47	EVFSQSSSL	20.000	103
	10	110	RVGGQNAVL	20.000	91
	11	218	IVGGNMSLL	20.000	55
	12	36	VAAQILSLL	12.000	104
	13	255	TAAHCVYDL	12.000	105
20	14	10	EAPFSFRSL	12.000	94
	15	138	YANVACAQL	12.000	106
	16	195	CASGHVVTL	12.000	107
	17	215	SSRIVGGNM	10.00	108
	18	298	KPKRLGNDI	8.000	109
25	19	313	GPLTFNEMI	8.000	110
	20	108	CVRVGGQNA	5.000	111
HLA B8					
	1	294	HSKYKPKRL	80.000	112
	2	373	RDVYGGIIS	16.000	100

	3	177	LPDDKV TAL	4.800	102
	4	265	LPKSWTIQV	2.400	113
	5	88	ELITRCDGV	2.400	114
	6	298	KPKRLGNDI	2.000	109
5	7	81	RSSF KCI EL	2.000	89
	8	375	VYGGIISPS	2.000	97
	9	79	RCRSSFKCI	2.000	115
	10	10	EAPFSFRSL	1.600	94
	11	215	SSRIVGGNM	1.000	108
10	12	36	VAAQILSLL	0.800	104
	13	255	TAAHC VYDL	0.800	116
	14	381	SPSMLCAGY	0.800	98
	15	195	CASGHVVTL	0.800	107
	16	362	VPLISNKIC	0.800	99
15	17	138	YANVACAQL	0.800	106
	18	207	ACGHRRGYS	0.400	117
	19	154	SDNLRVSSL	0.400	118
	20	47	EVFSQSSSL	0.400	103
HLA B2702					
20	1	300	KRLGNDIAL	180.000	119
	2	435	TRVTSFLDW	100.000	120
	3	376	YGGIISPSM	100.000	121
	4	410	RRLWKLVGA	60.000	122
	5	210	HRRGYSSRI	60.000	123
25	6	227	SQWPWQASL	30.000	49
	7	109	VRVGGQNAV	20.000	124
	8	191	VREGCASGH	20.000	125
	9	78	YRCRSSFKC	20.000	126
	10	113	GQNAVLQVF	20.000	127

	1 1	9 1	TRCDGVSDC	20.000	128
	1 2	3 8	AQILSLLPF	20.000	129
	1 3	2 1 1	RRGYSSRIV	18.000	130
	1 4	2 1 6	SRIVGGNMS	10.000	131
5	1 5	1 1 8	LQVFTAASW	10.000	132
	1 6	3 7 0	CNHRDVYGG	10.000	133
	1 7	3 9 3	GVDSCQGDS	10.000	134
	1 8	2 3 5	LQFQGYHLC	10.000	135
	1 9	2 7 1	IQVGLVSL	6.000	57
10	2 0	4 0 8	CQERRLWKL	6.000	95
	HLA B4403				
	1	4 2 7	AEVNKPGVY	90.000	136
	2	1 6 2	LEGQFREEF	40.000	137
	3	9	VEAPFSFRS	24.000	138
15	4	3 1 8	NEMIQPVCL	12.000	139
	5	2 5 6	AAHCVYDLY	9.000	140
	6	9 8	DCKDGEDEY	9.000	67
	7	4 6	FEVFSQSSS	8.000	141
	8	3 8	AQILSLLPF	7.500	129
20	9	6 4	LAIGLGIHF	7.500	142
	1 0	1 9 2	REGCASGHV	6.000	143
	1 1	3 3 0	EENFPDGKV	6.000	144
	1 2	1 8 2	VTALHHSVY	6.000	145
	1 3	4 0 8	QERRLWKLV	6.000	146
25	1 4	2 0 6	TACGHRRGY	4.500	147
	1 5	5	DPPAVEAPF	4.500	148
	1 6	2 6 1	YDLYLPKSW	4.500	149
	1 7	3 3	ADAVAAQIL	4.500	150
	1 8	1 6 8	EEFVSIDHL	4.000	151

19	304	NDIALMKLA	3.750	152
20	104	DEYRCVRVG	3.600	153

5 Conclusion

In this study, a serine protease was identified by means of a PCR based strategy. By Northern blot, the largest transcript for this gene is approximately 2.4 kb, and it is found to be expressed at high levels in ovarian tumors while found at minimal levels in all other tissues examined. The full-length cDNA encoding a novel multi-domain, cell-surface serine protease was cloned, named TADG-12. The 454 amino acid protein contains a cytoplasmic domain, a type II transmembrane domain, an LDLR-A domain, an SRCR domain and a serine protease domain. Using a semi-quantitative PCR analysis, it was shown that TADG-12 is overexpressed in a majority of tumors studied. Immunohistochemical staining corroborates that in some cases this protein is localized to the cell-surface of tumor cells and this suggests that TADG-12 has some extracellular proteolytic functions. Interestingly, TADG-12 also has a variant splicing form that is present in 35% of the tumors studied. This variant mRNA would lead to a truncated protein that may provide a unique peptide sequence on the surface of tumor cells.

This protein contains two extracellular domains which might confer unusual properties to this multidomain molecule. Although the precise role of LDLR-A function with regard to proteases remains unclear, this domain certainly has the capacity to bind calcium and other positively charged ligands [21,22]. This may play an important role in the regulation of the protease or

subsequent internalization of the molecule. The SRCR domain was originally identified within the macrophage scavenger receptor and functionally described to bind lipoproteins. Not only are SRCR domains capable of binding lipoproteins, but they may also bind to molecules as diverse as polynucleotides [23]. More recent studies have identified members of this domain family in proteins with functions that vary from proteases to cell adhesion molecules involved in maturation of the immune system [24]. In addition, TADG-12, like TMPRSS2 has only four of six cysteine residues conserved within its SRCR domain. This difference may allow for different structural features of these domains that confer unusual ligand binding properties. At this time, only the function of the CD6 encoded SRCR is well documented. In the case of CD6, the SRCR domain binds to the cell adhesion molecule ALCAM [23]. This mediation of cell adhesion is a useful starting point for future research on newly identified SRCR domains, however, the possibility of multiple functions for this domain can not be overlooked. SRCR domains are certainly capable of cell adhesion type interactions, but their capacity to bind other types of ligands should be considered.

At this time, the precise role of TADG-12 remains unclear. Substrates have not been identified for the protease domain, nor have ligands been identified for the extracellular LDLR-A and SRCR domains. Figure 8 presents a working model of TADG-12 with the information disclosed in the present invention. Two transcripts are produced which lead to the production of either TADG-12 or the truncated TADG-12V proteins. Either of these proteins is potentially targeted to the cell surface. TADG-12 is capable of becoming an activated serine protease while TADG-

12V is a truncated protein product that if at the cell surface may represent a tumor specific epitope.

The problem with treatment of ovarian cancer today remains the inability to diagnose the disease at an early stage.

5 Identifying genes that are expressed early in the disease process such as proteases that are essential for tumor cell growth [26] is an important step toward improving treatment. With this knowledge, it may be possible to design assays to detect the highly expressed genes such as the TADG-12 protease described

10 here or previously described proteases to diagnose these cancers at an earlier stage. Panels of markers may also provide prognostic information and could lead to therapeutic strategies for individual patients. Alternatively, inhibition of enzymes such as proteases may be an effective means for slowing progression of ovarian

15 cancer and improving the quality of patient life. Other features of TADG-12 and TADG-12V must be considered important to future research too. The extracellular ligand binding domains are natural targets for drug delivery systems. The aberrant peptide associated with the TADG-12V protein may provide an excellent

20 target drug delivery or for immune stimulation.

The following references were cited herein.

1. Duffy, M.J., Clin. Exp. Metastasis, 10: 145-155, 1992.
2. Monsky, W.L., et al., Cancer Res., 53: 3159-3164, 1993.
3. Powell, W.C., et al., Cancer Res., 53: 417-422, 1993.
- 25 4. Neurath, H. The Diversity of Proteolytic Enzymes. In: R.J. Beynon and J.S. Bond (eds.), pp. 1- 13, Proteolytic enzymes, Oxford: IRL Press, 1989.
5. Liotta, L.A., et al., Cell, 64: 327-336, 1991.

6. Tryggvason, K., et al., *Biochem. Biophys. Acta.*, 907: 191-217, 1987.
7. McCormack, R.T., et al., *Urology*, 45:729-744, 1995.
8. Landis, S.H., et al., *CA Cancer J. Clin.*, 48: 6-29, 1998.
- 5 9. Tanimoto, H., et al., *Cancer Res.*, 57: 2884-2887, 1997.
10. Tanimoto, H., et al., *Cancer*, 86: 2074-2082, 1999.
11. Underwood, L.J., et al., *Cancer Res.*, 59:4435-4439, 1999.
12. Tanimoto, et al., Increased Expression of Protease M in Ovarian Tumors. *Tumor Biology*, In Press, 2000.
- 10 13. Tanimoto, H., et al., *Proc. Of the Amer. Assoc. for Canc. Research* 39:648, 1998.
14. Tanimoto, H., et al., *Tumor Biology*, 20: 88-98, 1999.
15. Maniatis, T., Fritsch, E.F. & Sambrook, J. *Molecular Cloning*, p. 309-361 Cold Spring Harbor Laboratory, New York, 1982.
- 15 16. Shigemasa, K., et al., *J. Soc. Gynecol. Invest.*, 4:95-102, 1997.
17. Leytus, S.P., et al., *Biochemistry*, 27: 1067-1074, 1988.
18. Paoloni-Giacobino, A., et al., *Genomics*, 44: 309-320, 1997.
19. Sudhof, T.C., et al., *Science*, 228: 815-822, 1985.
20. Daly, N., et al., *Proc. Natl. Acad. Sci. USA* 92: 6334-6338, 1995.
- 20 21. Mahley, R.W., *Science* 240: 622-630, 1988.
22. Van Driel, I.R., et al., *J. Biol. Chem.* 262: 17443-17449, 1987.
23. Freeman, M., et al., *Proc. Natl. Acad. Sci. USA* 87: 8810-8814, 1990.
24. Aruffo, A., et al., *Immunology Today* 18(10): 498-504, 1997.
- 25 25. Rawlings, N.D., and Barrett, A.J., *Methods Enzymology* 244: 19-61, 1994.
26. Torres-Rosado, A., et al., *Proc. Natl. Acad. Sci. USA*, 90: 7181-7185, 1993.

Any patents or publications mentioned in this specification are indicative of the levels of those skilled in the art to which the invention pertains. These patents and publications are herein incorporated by reference to the same extent as if each
5 individual publication was specifically and individually indicated to be incorporated by reference.

One skilled in the art will readily appreciate that the present invention is well adapted to carry out the objects and obtain the ends and advantages mentioned, as well as those
10 inherent therein. The present examples along with the methods, procedures, treatments, molecules, and specific compounds described herein are presently representative of preferred embodiments, are exemplary, and are not intended as limitations on the scope of the invention. Changes therein and other uses will
15 occur to those skilled in the art which are encompassed within the spirit of the invention as defined by the scope of the claims.

WHAT IS CLAIMED IS:

1. A DNA fragment encoding Tumor Associated Differentially-Expressed Gene-12 (TADG-12) protein selected from
5 the group consisting of:

(a) an isolated DNA fragment which encodes a TADG-12 protein;

(b) an isolated DNA fragment which hybridizes to isolated DNA fragment of (a) above and which encodes a TADG-12
10 protein; and

(c) an isolated DNA fragment differing from the isolated DNA fragments of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-12 protein.
15

2. The DNA fragment of claim 1, wherein said DNA fragment has the sequence selected from the group consisting of SEQ ID No. 1 and SEQ ID No. 3.

20 3. The DNA fragment of claim 1, wherein said TADG-12 protein has the amino acid sequence selected from the group consisting of SEQ ID No. 2 and SEQ ID No. 4.

25 4. A vector comprising the DNA fragment of claim 1 and regulatory elements necessary for expression of the DNA in a cell.

5. The vector of claim 4, wherein said DNA fragment encodes a TADG-12 protein having the amino acid

sequence selected from the group consisting of SEQ ID No. 2 and SEQ ID No. 4.

6. A host cell transfected with the vector of claim 4,
5 said vector expressing a TADG-12 protein.

7. The host cell of claim 6, wherein said cell is selected from the group consisting of a bacterial cell, a mammalian cell, a plant cell and an insect cell.

10

8. The host cell of claim 7, wherein said bacterial cell is *E. coli*.

9. An antisense oligonucleotide directed against the
15 DNA fragment of claim 1.

10. An isolated and purified TADG-12 protein coded for by DNA selected from the group consisting of:

- 20 (a) isolated DNA which encodes a TADG-12 protein;
(b) isolated DNA which hybridizes to isolated DNA of (a) above and which encodes a TADG-12 protein; and
(c) isolated DNA differing from the isolated DNAs of (a) and (b) above in codon sequence due to the degeneracy of the genetic code, and which encodes a TADG-12 protein.

25

11. The isolated and purified TADG-12 protein of claim 10, wherein said TADG-12 protein has an amino acid sequence selected from the group consisting of SEQ ID No. 2 and SEQ ID No. 4.

12. A method for detecting expression of the TADG-12 protein of claim 10, comprising the steps of:

(a) contacting mRNA obtained from a cell with a labeled hybridization probe; and

5 (b) detecting hybridization of the probe with the mRNA.

13. An antibody directed against the TADG-12 protein of claim 10.

10

14. A method for diagnosing a cancer in an individual, comprising the steps of:

(a) obtaining a biological sample from said individual; and

15 (b) detecting a TADG-12 protein in said sample, wherein the presence of a TADG-12 protein in said sample is indicative of the presence of a cancer in said individual, wherein the absence of a TADG-12 protein in said sample is indicative of the absence of a cancer in said individual.

20

15. The method of claim 14, wherein said biological sample is selected from the group consisting of blood, urine, saliva, tears, interstitial fluid, ascites fluid, tumor tissue biopsy and circulating tumor cells.

25

16. The method of claim 14, wherein said detection of a TADG-12 protein is by means selected from the group consisting of Northern blot, Western blot, PCR, dot blot, ELIZA

sandwich assay, radioimmunoassay, DNA array chips and flow cytometry.

17. The method of claim 14, wherein said cancer is
5 selected from the group consisting of ovarian cancer, breast cancer, lung cancer, colon cancer, prostate cancer and other cancers in which TADG-12 is overexpressed.

18. A method for detecting malignant hyperplasia in
10 a biological sample, comprising the steps of:

- (a) isolating mRNA from said sample; and
 - (b) detecting TADG-12 mRNA in said sample,
- wherein the presence of said TADG-12 mRNA in said sample is indicative of the presence of malignant hyperplasia, wherein the
15 absence of said TADG-12 mRNA in said sample is indicative of the absence of malignant hyperplasia.

19. The method of claim 18, further comprising the step of comparing said TADG-12 mRNA to reference information,
20 wherein said comparison provides a diagnosis of said malignant hyperplasia.

20. The method of claim 18, further comprising the step of comparing said TADG-12 mRNA to reference information,
25 wherein said comparison determines a treatment of said malignant hyperplasia.

21. The method of claim 18, wherein said detection of TADG-12 mRNA is by PCR amplification.

22. The method of claim 21, wherein said PCR amplification uses primers selected from the group consisting of SEQ ID Nos. 28-31.

5 23. The method of claim 18, wherein said biological sample is selected from the group consisting of blood, urine, saliva, tears, interstitial fluid, ascites fluid, tumor tissue biopsy and circulating tumor cells.

10 24. A method for detecting malignant hyperplasia in a biological sample, comprising the steps of:

(a) isolating protein from said sample; and

(b) detecting a TADG-12 protein in said sample, wherein the presence of a TADG-12 protein in said sample is
15 indicative of the presence of malignant hyperplasia, wherein the absence of a TADG-12 protein in said sample is indicative of the absence of malignant hyperplasia.

20 25. The method of claim 24, further comprising the step of comparing said TADG-12 protein to reference information, wherein said comparison provides a diagnosis of said malignant hyperplasia.

25 26. The method of claim 24, further comprising the step of comparing said TADG-12 protein to reference information, wherein said comparison determines a treatment of said malignant hyperplasia.

27. The method of claim 24, wherein said detection is by immunoaffinity to an antibody, wherein said antibody is directed against a TADG-12 protein.

5 28. The method of claim 24, wherein said biological sample is selected from the group consisting of blood, urine, saliva, tears, interstitial fluid, ascites fluid, tumor tissue biopsy and circulating tumor cells.

10 29. A method of inhibiting expression of endogenous TADG-12 mRNA in a cell, comprising the step of:

 introducing a vector into a cell, wherein said vector comprises a DNA fragment of TADG-12 in opposite orientation operably linked to elements necessary for expression, wherein
15 expression of said vector in said cell produces TADG-12 antisense mRNA, wherein said TADG-12 antisense mRNA hybridizes to endogenous TADG-12 mRNA, thereby inhibiting expression of endogenous TADG-12 mRNA in said cell.

20 30. A method of inhibiting expression of a TADG-12 protein in a cell, comprising the step of:

 introducing an antibody into a cell, wherein said antibody is directed against a TADG-12 protein or fragment thereof, wherein binding of said antibody to said TADG-12 protein
25 or fragment thereof inhibits expression of said TADG-12 protein.

 31. A method of targeted therapy to an individual, comprising the step of:

administering a compound to an individual, wherein said compound has a targeting moiety and a therapeutic moiety, wherein said targeting moiety is specific for a TADG-12 protein.

5 32. The method of claim 31, wherein said targeting moiety is selected from the group consisting of an antibody directed against a TADG-12 protein and a ligand or ligand binding domain that binds a TADG-12 protein.

10 33. The method of claim 32, wherein said TADG-12 protein has an amino acid sequence selected from the group consisting of SEQ ID No. 2 and SEQ ID No. 4.

15 34. The method of claim 31, wherein said therapeutic moiety is selected from the group consisting of a radioisotope, a toxin, a chemotherapeutic agent, an immune stimulant and a cytotoxic agent.

20 35. The method of claim 31, wherein said individual suffers from a disease selected from the group consisting of ovarian cancer, lung cancer, prostate cancer, colon cancer and other cancers in which TADG-12 is overexpressed.

25 36. A method of vaccinating an individual against TADG-12, comprising the step of inoculating the individual with a TADG-12 protein or fragment thereof, wherein said TADG-12 protein or fragment thereof lacks TADG-12 activity, wherein said inoculation with said TADG-12 protein or fragment thereof elicits

an immune response in said individual, thereby vaccinating said individual against TADG-12.

37. The method of claim 36, wherein said individual
5 has a cancer, is suspected of having a cancer or is at risk of getting a cancer.

38. The method of claim 36, wherein said TADG-12 protein has an amino acid sequence selected from the group consisting of SEQ ID No. 2 and SEQ ID No. 4.

10

39. The method of claim 36, wherein said TADG-12 fragment has a sequence shown in SEQ ID No. 8.

40. The method of claim 36, wherein said TADG-12
15 fragment is a 9-residue fragment selected from the group consisting of SEQ ID Nos. 35, 36, 55, 56, 83, 84, 97, 98, 119, 120, 122, 123 and 136.

41. An immunogenic composition, comprising an
20 immunogenic fragment of a TADG-12 protein and an appropriate adjuvant.

42. The immunogenic composition of claim 41, wherein said immunogenic fragment of a TADG-12 protein has a sequence shown in SEQ ID No. 8.

25

43. The immunogenic composition of claim 41, wherein said immunogenic fragment of a TADG-12 protein is a 9-residue fragment selected from the group consisting of SEQ ID Nos. 35, 36, 55, 56, 83, 84, 97, 98, 119, 120, 122, 123 and 136.

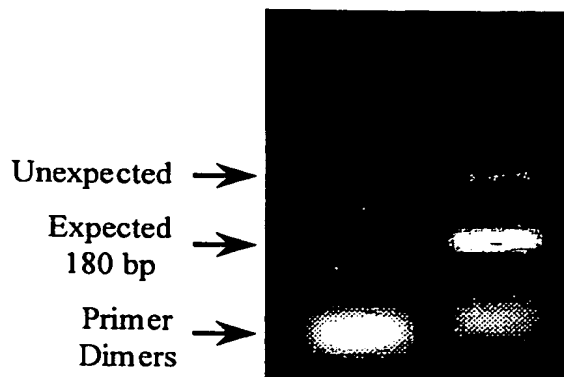


FIG. 1A

TADG12

↓

```

1  TGGGTGGTGACGGCGGCGCACTGTGTTTATGACTTGTACCTCCCCAAGTCATGGACCATC
   W V V T A A (H) C V Y D L Y L P K S W T I

61  CAGGTGGGTCTAGTTTCCCTGTTGGACAATCCAGCCCCATCCCACTTGGTGGAGAAGATT
   Q V G L V S L L D N P A P S H L V E K I
                                     (SEQ ID NO. 5)
121 GTCTACCACAGCAAGTACAAGCCAAAGAGGCTGGGCAACGACATCGCCCTCCTA
   V Y H S K Y K P K R L G N (D) I A L L
                                     (SEQ ID NO. 6)
  
```

TADG12-V

↓

```

1  GGGTGGTGACGGCGGCGCACTGTGTTTATGAGATTGTAGCTCCTAGAGAAAGGGCAGACA
   V V T A A H C V Y E I V A P R E R A D R

61  GAAGAGGAAGGAAGCTCCTGTGCTGGAGGAAACCCACAAAATGAAAGGACCTAGACCTT
   R G R K L L C W R K P T K M K G P R P S

121 CCCATAGCTAATTCCAGTGGACCATGTTATGGCAGATACAGGCTTGTACCTCCCCAAGTC
   H S * (SEQ ID NO. 8)

181 ATGGACCATCCAGGTGGGTCTAGTTTCCCTGTTGGACAATCCAGCCCCATCCCACTTGGT
241 GGAGAAGATTGTCTACCACAGCAAGTACAAGCCAAAGAGGCTGGGCAACGACATCGCCCT
301 CCTAATCACTAGTGCGGCCGCCTGCAGG (SEQ ID NO. 7)
  
```

FIG. 1B

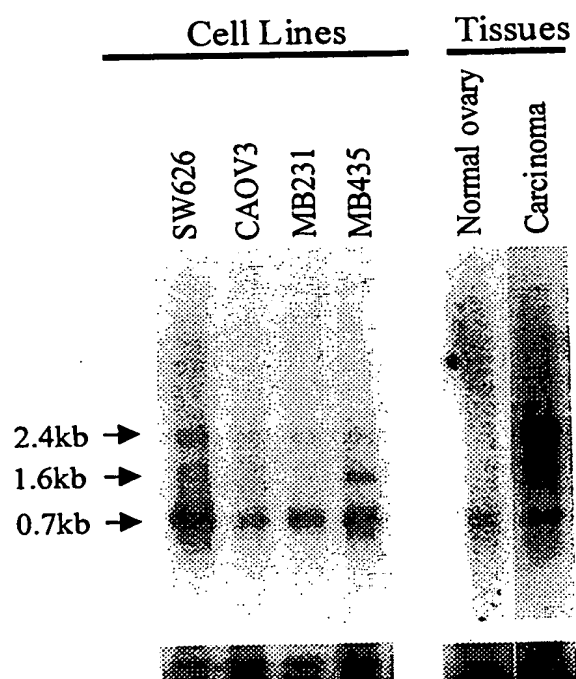


FIG. 2

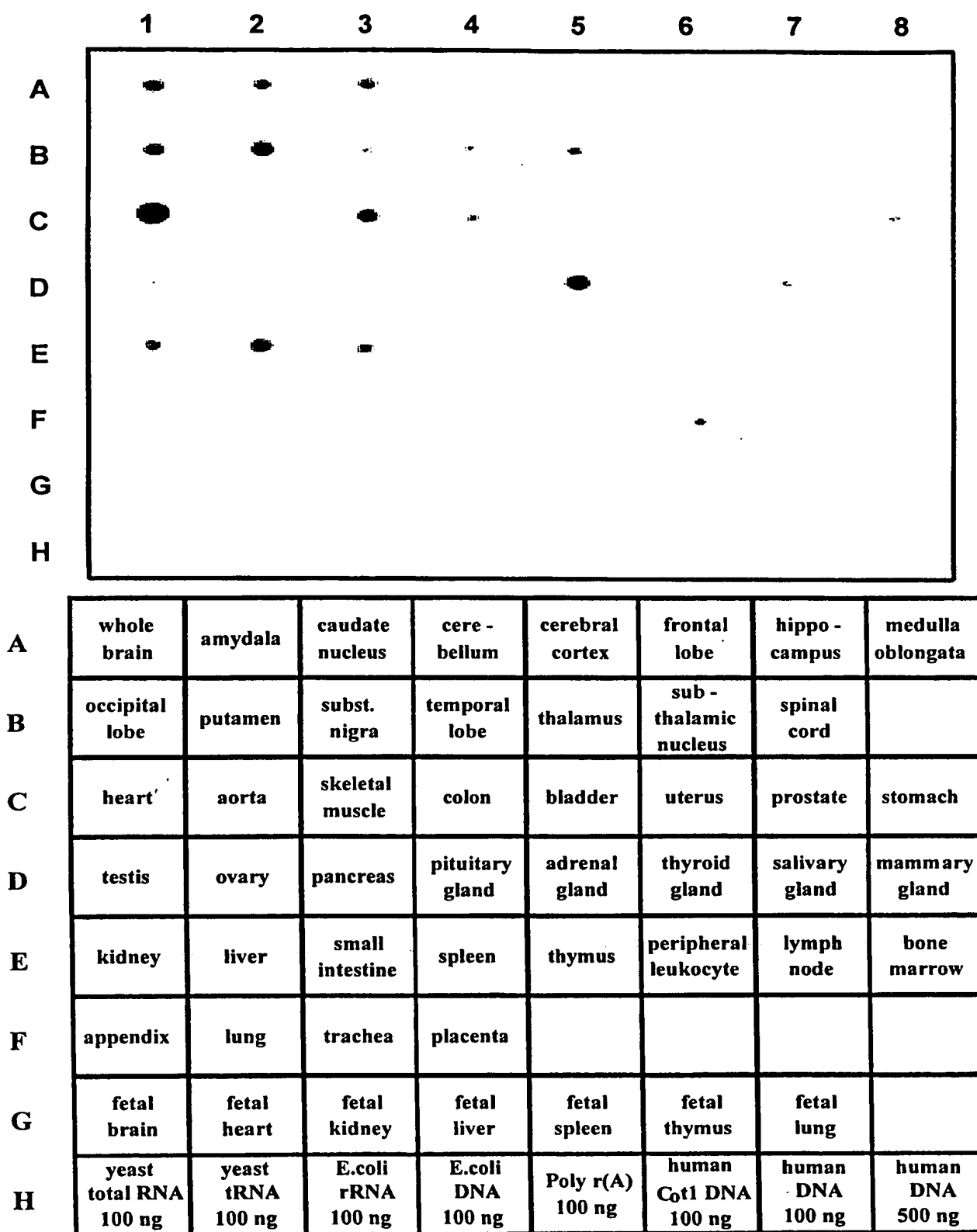


FIG. 3

1 CGGGAAAGGGCTGTGTTTATGGGAAGCCAGTAACACTGTGGCCTACTATCTCTTCCGTGG
61 TGCCATCTACATTTTGGGACTCGGGAATTATGAGGTAGAGGTGGAGGCGGAGCCGGATG
121 TCAGAGGTCTGAAATAGTCACCATGGGGGAAATGATCCGCCTGCTGTTGAAGCCCCCT
M G E N D P P A V E A P F 13
181 TCTCATTCCGATCGCTTTTTGGCCTTGATGATTTGAAAATAAGTCCTGTTGCACCAGATG
S F R S L F G L D L K I S P V A P D A 33
241 CAGATGCTGTTGCTGCACAGATCCTGTCACTGCTGCCATTGGAAGTTTTTCCCAATCAT
D A V A A Q I L S L L P F E V F S Q S S 53
301 CGTCATTGGGGATCATTGCATTGATATTAGCACTGGCCATTGGTCTGGGCATCCACTTCG
S L G I I A L I L A L A I G L G I H F D 73
361 ACTGCTCAGGGAAGTACAGATGTCGCTCATCCTTTAAGTGTATCGAGCTGATAACTCGAT
C S G K Y R C R S S F K C I E L I T R C 93
421 GTGACGGAGTCTCGGATTGCAAAGACGGGGAGGAGTACCGCTGTGTCCGGGTGGGTG
D G V S D C K D G E D E Y R C V R V G G 113
481 GTCAGAATGCCGTGCTCCAGGTGTTACAGCTGCTTCGTGGAAGACCATGTGCTCCGATG
O N A V L Q V F T A A S W K T M C S D D 133
541 ACTGGAAGGGTCACTACGCAAATGTTGCCTGTGCCAACTGGGTTTCCCAAGCTATGTGA
W K G H Y A N V A C A Q L G F P S Y V S 153
601 GTTCAGATAACCTCAGAGTGAGCTCGCTGGAGGGGAGTTCCGGGAGGAGTTTGTGTCCA
S D N L R V S S L E G Q F R E E F V S I 173
661 TCGATCACCTCTTGCCAGATGACAAGGTGACTGCATTACACCACTCAGTATATGTGAGGG
D H L L P D D K V T A L H H S V Y V R E 193
721 AGGGATGTGCTCTGGCCACGTGGTTACCTTGCACTGCACAGCCTGTGGTTCATAGAAGGG
G C A S G H V V T L Q C T A C G H R R G 213
781 GCTACAGCTCACGCATCGTGGGTGGAACATGTCCTTGCTCTCGCAGTGGCCCTGGCAGG
Y S S R I V G G N M S L L S Q W P W Q A 233
841 CCAGCCTTCAGTTCAGGGGCTACCACCTGTGCGGGGGCTCTGTCATCAGCCCCCTGTGGA
S L Q F Q G Y H L C G G S V I T P L W I 253
901 TCATCACTGCTGCACACTGTGTTTATGACTTGTAACCTCCCCAAGTCATGGACCATCCAGG
I T A A H C V Y D L Y L P K S W T I Q V 273
961 TGGGTCTAGTTTCCCTGTTGGACAATCCAGCCCCATCCCACCTGGTGGAGAAGATTGTCT
G L V S L L D N P A P S H L V E K I V Y 293
1021 ACCACAGCAAGTACAAGCCAAAGAGGCTGGGCAATGACATCGCCCTTATGAAGCTGGCCG
H S K Y K P K R L G N D I A L M K L A G 313
1081 GGCCACTCACGTTCAATGAATGATCCAGCCTGTGTGCCTGCCCAACTCGAAGAGAAGT
P L T F N E M I Q P V C L P N S E E N F 333
1141 TCCCCGATGGAAAAGTGTGCTGGACGTCAGGATGGGGGGCCACAGAGGATGGAGGTGACG
P D G K V C W T S G W G A T E D G G D A 353
1201 CCTCCCCTGTCCTGAACCACGCGGCCGTCCCTTTGATTTCCAACAAGATCTGCAACCACA
S P V L N H A A V P L I S N K I C N H R 373
1261 GGGACGTGTACGGTGGCATCATCTCCCCCTCCATGCTCTGCGCGGGCTACCTGACGGGTG
D V Y G G I I S P S M L C A G Y L T G G 393
1321 GCGTGGACAGCTGCCAGGGGGACAGCGGGGGGGCCCTGGTGTGTCAAGAGAGGAGGCTGT
V D S C Q G D S G G P L V C Q E R R L W 413
1381 GGAAGTTAGTGGGAGCGACCACTTTGGCATCGGCTGCGCAGAGGTGAACAAGCCTGGGG
K L V G A T S F G I G C A E V N K P G V 433
1441 TGTACACCGTGTACCTCCTTCTGGAAGTGGATCCACGAGCAGATGGAGAGAGACCTAA
Y T R V T S F L D W I H E Q M E R D L K 453
1501 AAACCTGAAGAGGAAGGGGACAAGTAGCCACCTGAGTTCCTGAGGTGATGAAGACAGCCC
T * (SEQ ID NO. 2) 454
1561 GATCCTCCCCTGGACTCCCCTGTAGGAACCTGCACACGAGCAGACACCCTTGGAGCTCTG
1621 AGTTCGGCACCAGTAGCGGGCCGAAAGAGGCACCCTTCCATCTGATTCCAGCACAACC
1681 TTCAAGCTGCTTTTTTGTGTTTTTGTGTTTTTGTAGGTGGAGTCTCGCTCTGTTGCCAGGCT
1741 GGAGTGCAGTGGCGAAATACCCTGCTCACTGCAGCCTCCGCTTCCCTGTTTCAAGCGATT
1801 CTCTTGCCTCAGCTTCCCCAGTAGCTGGGACCACAGGTGCCCGCCACCACACCACTAA
1861 TTTTTGTATTTTGTAGAGACAGGGTTTACCATTGTTGGCCAGGCTGCTCTCAACCC
1921 TGACCTCAAATGATGTGCTGCTTACGCTCCACAGTGTGCTGGGATTACAGGCATGGGCC
1981 ACCACGCTAGCCTACGCTCCTTTCTGATCTTCACTAAGAACAAAAGAAGCAGCAACTT
2041 GCAAGGGCGGCCCTTTCCCACTGGTCCATCTGGTTTTCTCTCCAGGGTCTTGCAAAATTCC
2101 TGACGAGATAAGCAGTTATGTGACCTACGTCGAAAGCCACCAACAGCCACTCAGAAAAG
2161 ACGCACCAGCCCAGAAGTGCAGAAGTGCAGTCACTGCACGTTTTTCATCTTTAGGGACCAG
2221 AACCAAACCCACCCTTTCTACTTCAAGACTTATTTTACATGTGGGGAGGTTAATCTAG
2281 GAATGACTCGTTTAAAGGCCTATTTTCATGATTTCTTTGTAGCATTTGGTGCTTGACGTAT
2341 TATTGTCCTTTGATTCCAAATAATATGTTTCTTCCCTCAAAAAAAAAAAAAAAAAAAAA
2401 AAAAAAAAAAAAAA (SEQ ID NO. 1)

FIG. 4

Comp8	CEG..FVC	AQTGRCVNR	LLCNGDNDCG	DQSDEAN.C	(SEQ ID NO. 9)
Matr	CPG.QFTC	.RTGRCIRKE	LRCDGWADCT	DHSDELN.C	(SEQ ID NO. 10)
Gp300-1	CQQGYFKC	QSEGQCIPSS	WVCDQDQDCD	DGSDERQDC	(SEQ ID NO. 11)
Gp300-2	CSSHQITC	.SNGQCIPSE	YRCDHVRCDCP	DGADE.NDC	(SEQ ID NO. 12)
TADG12	CSGK.YRC	RSSFKCIELI	TRCDGVSDCK	DGEDEYR.C	(SEQ ID NO. 13)
Tmprss2	CSNSGIEC	DSSGTCINPS	NWCDGVSHCP	GGEDENR.C	(SEQ ID NO. 14)
Cons	C	C	C	C	DE C

FIG. 5A

BovEntk	VRLVGGSGPH	EGRVEI.FHE	GQWGTVCDDR	WELRGGLVVC	RSLGYKGVQS
MacSR	VRLVGGSGPH	EGRVEI.LHS	GQWGTICDDR	WEVRVGQVVC	RSLGYPGVQA
TADG12	VRVGG...QN	AVLQVFTA..	ASWKTMCSD	WKGHYANVAC	AQLGFP.SYV
Tmprss2	VRLYG...PN	FILQMYSSQR	KSWHPVCQDD	WNENYGRAAC	RDMGYKNNFY
HumEntk	VREFNGTTNN	NGLVRFRIQ.	SIWHTACAEN	WTTQISNDVC	QLLGLGSG..
Cons	VR		W C	W	C

BovEntk	VHKRAYFGKG	TGPIWLNEVF	CFGK..ESSI	EECRIRQWGV	R.ACSHDEDA
MacSR	VHKAHFGQG	TGPIWLNEVF	CFGR..ESSI	EECKIRQWGT	R.ACSHSEDA
TADG12	SSDNLRVSSL	EGQFREEFVS	I.DHLLPDDK	VTALHHSVYV	REGCASGHVV
Tmprss2	SSQGIVDDSG	STSFMKLNTS	A.GNV...DI	YKKLYHS...	.DACSSKAVV
HumEntk	NSSKPIFSTD	GGPFVKLNTA	PDGHLILTPS	QQ.....	...CLQDSL
Cons					C

BovEntk	GVTCT	(SEQ ID NO. 15)
MacSR	GVTCT	(SEQ ID NO. 16)
TADG12	TLQCT	(SEQ ID NO. 17)
Tmprss2	SLRCL	(SEQ ID NO. 18)
HumEntk	RLQC.	(SEQ ID NO. 19)
Cons	C	

FIG. 5B

```

ProM  LWVLTAAHCK .....KPNL QVFLGKHNLR QRESSQEQSS VVRAVIHPDY
Try1   QWVVSAGHCY .....KSRI QVRLGEHNIE VLEGNEQFIN AAKIIRHPQY
Kal    QWVLTAAHCF D.GLPLQDVW RIYSGILNLS DITKDTFESQ IKEIIHQNY
TADG12 LWIITAACHV .YDLYLPKSW TIQVGLV..S LLDNPAPSHL VEKIVYHSKY
Tmprss2 EWIVTAACHV EKPLNNPWHW TAFAGILRQS FMYGA.GYQ VQKVISHPNY
Heps   DWVLTAACHF PERNRVLSRW RVFAGAVAQA SPHGLQLG.. VQAVVYHGGY
Cons   W   A HC                      G                      H   Y

ProM  .....DAAS HDQDIMLLRL ARPAKLSELI QPLPLERDCS ANT..TSCHI
Try1   .....DRKT LNNDIMLIK L SSRAVINARV STISLPTAPP ATG..TKCLI
Kal    .....KVSE GNHDIALIKL QAPLNYTEFQ KPICLPSKGD TSTIYTNCWV
TADG12 .....KPKR LGNDIALMKL AGPLTFNEMI QPVCLPNSEE NFPDGKVCWT
Tmprss2 .....DSKT KNNDIALMKL QKPLTFNDLV KPVCLPNPGM MLQPEQLCWI
Heps   LPFRDPNSEE NSNDIALVHL SSPLPLTEYI QPVCLPAAGQ ALVDGKICTV
Cons   DI L L                      L                      C

ProM  LGWGKTAD.. GDFPDTIQCA YIHLVSREEC EHA..YPGQI TQNMLCAGDE
Try1   SGWGNTASSG ADYPDELQCL DAPVLSQAKC EAS..YPGKI TSNMFCVGFL
Kal    TGWGFSKEK. GEIQNILQKV NIPLVTNEEC QKR.YQDYKI TORMVCAGYK
TADG12 SGWGAT.EDG GDASPVLNHA AVPLISNKIC NHRDVYGGII SPSMLCAGYL
Tmprss2 SGWGAT.EEK GKTSEVLNAA KVLLIETQRC NSRYVYDNLI TPAMICAGFL
Heps   TGWGNT.QYY GQQAGVLQEA RVPIISNDVC NGADFYGNQI KPKMFCAGYP
Cons   GWG                      C                      I      M C G

ProM  KYGKDSCQGD SGGPLVC (SEQ ID NO. 20)
Try1   EGGKDSCQGD SGGPVVC (SEQ ID NO. 21)
Kal    EGGKDACKGD SGGPLVC (SEQ ID NO. 22)
TADG12 TGGVDSCQGD SGGPLVC (SEQ ID NO. 23)
Tmprss2 QGNVDSCQGD SGGPLVT (SEQ ID NO. 24)
Heps   EGGIDACQGD SGGPFVC (SEQ ID NO. 25)
Cons   D C GD SGGP V

```

FIG. 5C

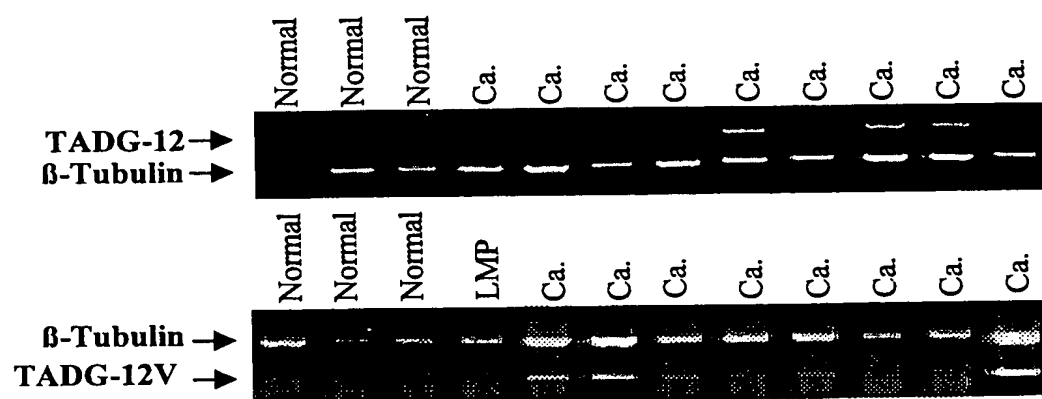


FIG. 6



FIG. 7A



FIG. 7B



FIG. 7C

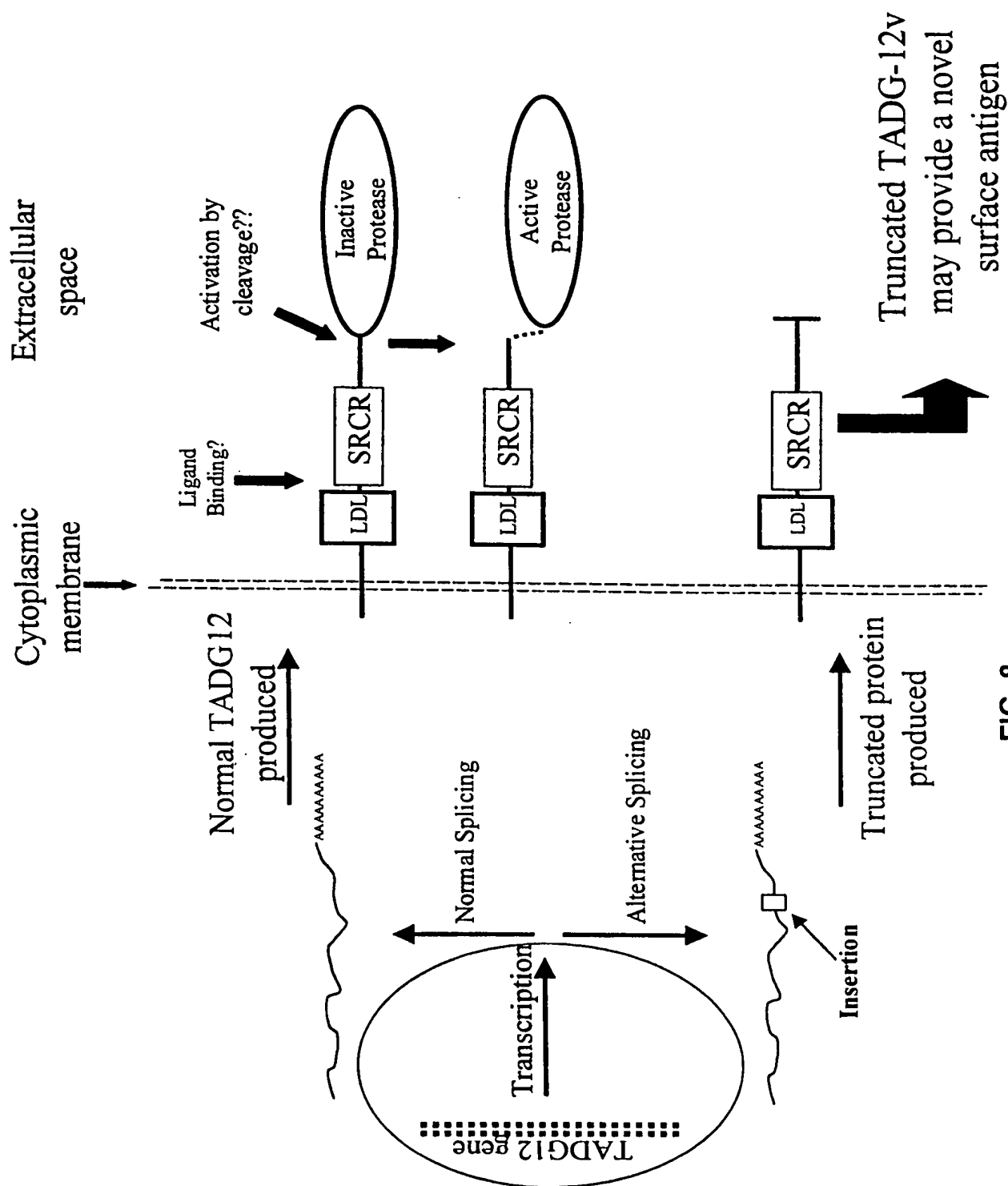


FIG. 8

SEQUENCE LISTING

<110> O'Brien, Timothy J.
 Underwood, Lowell J.
 <120> Transmembrane Serine Protease Overexpressed
 in Ovarian Carcinoma and Uses Thereof
 <130> D6192PCT
 <141> 2000-03-02
 <150> 09/261,416
 <151> 1999-03-03
 <160> 153

 <210> 1
 <211> 2413
 <212> DNA
 <213> *Homo sapiens*
 <220>
 <221> CDS
 <223> entire cDNA sequence of TADG-12 gene
 <400> 1

```

cgggaaaggg ctgtgtttat gggaagccag taacactgtg gcctactatc 50
tcttccgtgg tgccatctac atttttggga ctcggaatt atgaggtaga 100
ggtggaggcg gagccggatg tcagagggtc tgaaatagtc accatggggg 150
aaaatgatcc gctgtgtgtt gaagccccct tctcattccg atcgcttttt 200
ggccttgatg atttgaaaat aagtcctgtt gcaccagatg cagatgctgt 250
tgctgcacag atcctgtcac tgctgccatt tgaagttttt tcccaatcat 300
cgtcattggg gatcattgca ttgatattag cactggccat tggcttgggc 350
atccacttcg actgctcagg gaagtacaga tgtcgctcat cctttaagtg 400
tatcgagctg ataactcgat gtgacggagt ctcgattgc aaagacgggg 450
aggacgagta ccgctgtgtc cgggtgggtg gtcagaatgc cgtgctccag 500
gtgttcacag ctgcttcgtg gaagaccatg tgctccgatg actggaaggg 550
tcactacgca aatgttgcc tggcccaact gggtttccca agctatgtga 600
gttcagataa cctcagagtg agctcgctgg aggggcagtt ccgggaggag 650
tttgtgtcca tcgatcacct cttgccagat gacaaggatg ctgcattaca 700
ccactcagta tatgtgaggg agggatgtgc ctctggccac gtggttacct 750
tgcagtgcac agcctgtggt catagaaggg gctacagctc acgcatcgtg 800
ggtggaaaca tgccttgct ctgcagtg ccctggcagg ccagccttca 850
gttccagggc taccacctgt gcggggggtc tgtcatcacg cccctgtgga 900
tcatcactgc tgcacactgt gtttatgact tgtacctccc caagtcatgg 950
accatccagg tgggtctagt ttccctgttg gacaatccag ccccatccca 1000
cttggtggag aagattgtct accacagcaa gtacaagcca aagaggctgg 1050
gcaatgacat cgcccttatg aagctggccg ggccactcac gttcaatgaa 1100
atgatccagc ctgtgtgcct gcccaactct gaagagaact tccccgatgg 1150
aaaagtgtgc tggacgtcag gatggggggc cacagaggat ggaggtgacg 1200
cctcccctgt cctgaaccac gcggccgtcc ctttgatttc caacaagatc 1250
tgcaaccaca gggacgtgta cgggtggcatc atctccccct ccatgctctg 1300
cgcgggctac ctgacgggtg gcgtgaacag ctgccagggg gacagcgggg 1350
ggcccttggg gtgtcaagag aggaggctgt ggaagttagt gggagcgacc 1400
agctttggca tcggctgcgc agaggtgaac aagcctgggg tgtacaccg 1450
tgtcacctcc ttcttgact ggatccacga gcagatggag agagacctaa 1500
aaacctgaag aggaagggga caagtagcca cctgagttcc tgaggtgatg 1550
aagacagccc gatcctcccc tggactcccc tgtaggaacc tgcacacgag 1600
cagacaccct tggagctctg agttccggca ccagtagcgg gcccgaaaga 1650
ggcacccttc catctgattc cagcacaacc ttcaagctgc tttttgtttt 1700
ttgttttttt gaggtggagt ctcgctctgt tgcccaggct ggagtgcagt 1750

```

```

ggcgaaatac cctgctcact gcagcctccg cttccctggt tcaagcgatt 1800
ctcttgccctc agcttcccca gtagctggga ccacaggtgc ccgccaccac 1850
acccaactaa tttttgtatt tttagtagag acagggtttc accatgttgg 1900
ccaggctgct ctcaaaccct tgacctcaaa tgatgtgcct gcttcagcct 1950
cccacagtgc tgggattaca ggcatgggcc accacgccta gcctcacgct 2000
cctttctgat cttcactaag aacaaaagaa gcagcaactt gcaagggcgg 2050
cctttcccac tgggtccatct gggttttctt ccaggggtctt gcaaaaattcc 2100
tgacgagata agcagttatg tgacctcacg tgcaaagcca ccaacagcca 2150
ctcagaaaag acgcaccagc ccagaagtgc agaactgcag tcaactgcacg 2200
ttttcatctt tagggaccag aaccaaacc accctttcta cttccaagac 2250
ttattttcac atgtggggag gttaatctag gaatgactcg ttttaaggcct 2300
attttcatga tttctttgta gcatttggtg cttgacgtat tattgtcctt 2350
tgattccaaa taatatgttt cttccctca aaaaaaaaaa aaaaaaaaaa 2400
aaaaaaaaaa aaa 2413

```

```

<210>      2
<211>      454
<212>      PRT
<213>      Homo sapiens
<220>
<223>      complete amino acid sequence of TADG-12
              protein
<400>      2

```

```

Met Gly Glu Asn Asp Pro Pro Ala Val Glu Ala Pro Phe Ser Phe
      5      10
Arg Ser Leu Phe Gly Leu Asp Asp Leu Lys Ile Ser Pro Val Ala
      20      25
Pro Asp Ala Asp Ala Val Ala Ala Gln Ile Leu Ser Leu Leu Pro
      35      40
Phe Glu Val Phe Ser Gln Ser Ser Ser Leu Gly Ile Ile Ala Leu
      50      55
Ile Leu Ala Leu Ala Ile Gly Leu Gly Ile His Phe Asp Cys Ser
      65      70
Gly Lys Tyr Arg Cys Arg Ser Ser Phe Lys Cys Ile Glu Leu Ile
      80      85
Thr Arg Cys Asp Gly Val Ser Asp Cys Lys Asp Gly Glu Asp Glu
      95     100
Tyr Arg Cys Val Arg Val Gly Gly Gln Asn Ala Val Leu Gln Val
     110     115
Phe Thr Ala Ala Ser Trp Lys Thr Met Cys Ser Asp Asp Trp Lys
     125     130
Gly His Tyr Ala Asn Val Ala Cys Ala Gln Leu Gly Phe Pro Ser
     140     145
Tyr Val Ser Ser Asp Asn Leu Arg Val Ser Ser Leu Glu Gly Gln
     155     160
Phe Arg Glu Glu Phe Val Ser Ile Asp His Leu Leu Pro Asp Asp
     170     175
Lys Val Thr Ala Leu His His Ser Val Tyr Val Arg Glu Gly Cys
     185     190
Ala Ser Gly His Val Val Thr Leu Gln Cys Thr Ala Cys Gly His
     200     205
Arg Arg Gly Tyr Ser Ser Arg Ile Val Gly Gly Asn Met Ser Leu
     215     220
Leu Ser Gln Trp Pro Trp Gln Ala Ser Leu Gln Phe Gln Gly Tyr
     230     235

```


His	Leu	Cys	Gly	Gly	Ser	Val	Ile	Thr	Pro	Leu	Trp	Ile	Ile	Thr
				245					250					255
Ala	Ala	His	Cys	Val	Tyr	Asp	Leu	Tyr	Leu	Pro	Lys	Ser	Trp	Thr
				260					265					270
Ile	Gln	Val	Gly	Leu	Val	Ser	Leu	Leu	Asp	Asn	Pro	Ala	Pro	Ser
				275					280					285
His	Leu	Val	Glu	Lys	Ile	Val	Tyr	His	Ser	Lys	Tyr	Lys	Pro	Lys
				290					295					300
Arg	Leu	Gly	Asn	Asp	Ile	Ala	Leu	Met	Lys	Leu	Ala	Gly	Pro	Leu
				305					310					315
Thr	Phe	Asn	Glu	Met	Ile	Gln	Pro	Val	Cys	Leu	Pro	Asn	Ser	Glu
				320					325					330
Glu	Asn	Phe	Pro	Asp	Gly	Lys	Val	Cys	Trp	Thr	Ser	Gly	Trp	Gly
				335					340					345
Ala	Thr	Glu	Asp	Gly	Gly	Asp	Ala	Ser	Pro	Val	Leu	Asn	His	Ala
				350					355					360
Ala	Val	Pro	Leu	Ile	Ser	Asn	Lys	Ile	Cys	Asn	His	Arg	Asp	Val
				365					370					375
Tyr	Gly	Gly	Ile	Ile	Ser	Pro	Ser	Met	Leu	Cys	Ala	Gly	Tyr	Leu
				380					385					390
Thr	Gly	Gly	Val	Asp	Ser	Cys	Gln	Gly	Asp	Ser	Gly	Gly	Pro	Leu
				395					400					405
Val	Cys	Gln	Glu	Arg	Arg	Leu	Trp	Lys	Leu	Val	Gly	Ala	Thr	Ser
				410					415					420
Phe	Gly	Ile	Gly	Cys	Ala	Glu	Val	Asn	Lys	Pro	Gly	Val	Tyr	Thr
				425					430					435
Arg	Val	Thr	Ser	Phe	Leu	Asp	Trp	Ile	His	Glu	Gln	Met	Glu	Arg
				440					445					450
Asp	Leu	Lys	Thr											

<210> 3
 <211> 2544
 <212> DNA
 <213> *Homo sapiens*
 <220>
 <221> CDS
 <223> entire cDNA sequence of TADG-12 variant gene
 <400> 3

cgggaaaggg	ctgtgtttat	gggaagccag	taacactgtg	gcctactatc	50
tcttccgtgg	tgccatctac	atttttggga	ctcgggaatt	atgaggtaga	100
ggtggaggcg	gagccggatg	tcagaggtcc	tgaaatagtc	accatggggg	150
aaaatgatcc	gcctgctgtt	gaagccccct	tctcattccg	atcgcttttt	200
ggccttgatg	atttgaaaat	aagtcctgtt	gcaccagatg	cagatgctgt	250
tgctgcacag	atcctgtcac	tgctgccatt	tgaagttttt	tcccaatcat	300
cgtcattggg	gatcattgca	ttgatattag	cactggccat	tggctctgggc	350
atccacttcg	actgctcagg	gaagtacaga	tgctcgctcat	cctttaagtg	400
tatcgagctg	ataactcgat	gtgacggagt	ctcggattgc	aaagacgggg	450
aggacgagta	ccgctgtgtc	cgggtgggtg	gtcagaatgc	cgtgctccag	500
gtgttcacag	ctgcttcgtg	gaagaccatg	tgctccgatg	actggaaggg	550
tcactacgca	aatgttgcc	gtgcccact	gggtttccca	agctatgtaa	600
gttcagataa	cctcagagt	agctcgctgg	aggggcagtt	ccgggaggag	650
tttgtgtcca	tcgatcacct	cttgccagat	gacaagggtga	ctgcattaca	700
ccactcagta	tatgtgaggg	agggatgtgc	ctctggccac	gtggttacct	750
tgcagtgcac	agcctgtggt	catagaaggg	gctacagctc	acgcatcgtg	800

```

ggtggaaaca tgtccttget ctgcagtggt ccctggcagg ccagccttca 850
gttccagggc taccacctgt gcggggggctc tgtcatcacg cccctgtgga 900
tcatcactgc tgcacactgt gtttatgaga ttgtagctcc tagagaaagg 950
gcagacagaa gaggaaggaa gctcctgtgc tggaggaaac ccacaaaaat 1000
gaaaggacct agaccttccc atagctaatt ccagtggacc atgttatggc 1050
agatacagggc ttgtacctcc ccaagtcacg gaccatccag gtgggtctag 1100
tttccctggt ggacaatcca gccccatccc acttggtgga gaagattgtc 1150
taccacagca agtacaagcc aaagaggctg ggcaatgaca tcgcccttat 1200
gaagctggcc gggccactca cgttcaatga aatgatccag cctgtgtgcc 1250
tgcccaactc tgaagagaaac ttccccgatg gaaaagtgtg ctggacgtca 1300
ggatgggggg ccacagagga tggaggtgac gcctccccctg tcctgaacca 1350
cgcggccgct cctttgattt ccaacaagat ctgcaaccac agggacgtgt 1400
acggtggcat catctcccc cccatgctct gcgcgggcta cctgacgggt 1450
ggcgtggaca gctgccaggg ggacagcggg gggccccctg tgtgtcaaga 1500
caggaggctg tggaagttag tgggagcgac cagctttggc atcggtctgcg 1550
cagaggtgaa caagcctggg gtgtacaccc gtgtcacctc cttcctggac 1600
tggatccacg agcagatgga gagagaccta aaaacctgaa gaggaagggg 1650
acaagtagcc acctgagttc ctgaggtgat gaagacagcc cgatcctccc 1700
ctggactccc gtgtaggaac ctgcacacga gcagacaccc ttggagctct 1750
gagttccggc accagtagcg ggcccgaag aggcaccctt ccatctgatt 1800
ccagcacaac cttcaagctg ctttttgttt tttgtttttt tgaggtggag 1850
tctcgtctcg ttgccagggc tggagtgcag tggcgaaata ccctgctcac 1900
tgcagcctcc gcttccctgg ttcaagcgat tctcttgct cagcttcccc 1950
agtagctggg accacaggtg cccgccacca caccctaata atttttgtat 2000
ttttagtaga gacagggttt caccatgttg gccaggctgc tctcaaacc 2050
ctgacctcaa atgatgtgcc tgcttcagcc tcccacagtg ctgggattac 2100
aggcatgggc caccacgcct agcctcacgc tcctttctga tcttactaa 2150
gaacaaaaga agcagcaact tgcaaggcg gcctttccca ctggtccatc 2200
tggttttctc tccagggtct tgcaaaattc ctgacgagat aagcagttat 2250
gtgacctcac gtgcaaagcc accaacagcc actcagaaaa gacgcaccag 2300
cccagaagtg cagaactgca gtcactgcac gttttcatct ttagggacca 2350
gaaccaaacc caccctttct acttccaaga cttattttca catgtgggga 2400
ggttaatcta ggaatgactc gtttaaggcc tattttcatg atttctttgt 2450
agcatttggt gcttgacgta ttattgtcct ttgattccaa ataatatggt 2500
tccttccttc aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaa 2544

```

```

<210>      4
<211>      294
<212>      PRT
<213>      Homo sapiens
<220>
<223>      complete amino acid sequence of TADG-12
variant protein
<400>      4

```

```

Met Gly Glu Asn Asp Pro Pro Ala Val Glu Ala Pro Phe Ser Phe
      5      10
Arg Ser Leu Phe Gly Leu Asp Asp Leu Lys Ile Ser Pro Val Ala
      20      25
Pro Asp Ala Asp Ala Val Ala Ala Gln Ile Leu Ser Leu Leu Pro
      35      40
Phe Glu Val Phe Ser Gln Ser Ser Ser Leu Gly Ile Ile Ala Leu
      50      55
Ile Leu Ala Leu Ala Ile Gly Leu Gly Ile His Phe Asp Cys Ser
      65      70
Gly Lys Tyr Arg Cys Arg Ser Ser Phe Lys Cys Ile Glu Leu Ile

```

				80					85					90
Thr	Arg	Cys	Asp	Gly	Val	Ser	Asp	Cys	Lys	Asp	Gly	Glu	Asp	Glu
				95					100					105
Tyr	Arg	Cys	Val	Arg	Val	Gly	Gly	Gln	Asn	Ala	Val	Leu	Gln	Val
				110					115					120
Phe	Thr	Ala	Ala	Ser	Trp	Lys	Thr	Met	Cys	Ser	Asp	Asp	Trp	Lys
				125					130					135
Gly	His	Tyr	Ala	Asn	Val	Ala	Cys	Ala	Gln	Leu	Gly	Phe	Pro	Ser
				140					145					150
Tyr	Val	Ser	Ser	Asp	Asn	Leu	Arg	Val	Ser	Ser	Leu	Glu	Gly	Gln
				155					160					165
Phe	Arg	Glu	Glu	Phe	Val	Ser	Ile	Asp	His	Leu	Leu	Pro	Asp	Asp
				170					175					180
Lys	Val	Thr	Ala	Leu	His	His	Ser	Val	Tyr	Val	Arg	Glu	Gly	Cys
				185					190					195
Ala	Ser	Gly	His	Val	Val	Thr	Leu	Gln	Cys	Thr	Ala	Cys	Gly	His
				200					205					210
Arg	Arg	Gly	Tyr	Ser	Ser	Arg	Ile	Val	Gly	Gly	Asn	Met	Ser	Leu
				215					220					225
Leu	Ser	Gln	Trp	Pro	Trp	Gln	Ala	Ser	Leu	Gln	Phe	Gln	Gly	Tyr
				230					235					240
His	Leu	Cys	Gly	Gly	Ser	Val	Ile	Thr	Pro	Leu	Trp	Ile	Ile	Thr
				245					250					255
Ala	Ala	His	Cys	Val	Tyr	Glu	Ile	Val	Ala	Pro	Arg	Glu	Arg	Ala
				260					265					270
Asp	Arg	Arg	Gly	Arg	Lys	Leu	Leu	Cys	Trp	Arg	Lys	Pro	Thr	Lys
				275					280					285
Met	Lys	Gly	Pro	Arg	Pro	Ser	His	Ser						
				290										

<210> 5
 <211> 174
 <212> DNA
 <213> Artificial sequence
 <220>
 <223> nucleotide sequence of the subclone containing
 the 180 bp band from the PCR product for TADG-12
 <400> 5

tgggtggtga	cggcggcgca	ctgtgtttat	gacttgtacc	tccccaagtc	50
atggaccatc	caggtgggtc	tagtttcct	gttgacaat	ccagcccat	100
cccacttggt	ggagaagatt	gtctaccaca	gcaagtacaa	gccaaagagg	150
ctgggcaacg	acatgcacct	ccta			174

<210> 6
 <211> 58
 <212> PRT
 <213> Artificial sequence
 <220>
 <223> deduced amino acid sequence of the 180 bp band
 from the PCR product for TADG-12
 <400> 6

Trp	Val	Val	Thr	Ala	Ala	His	Cys	Val	Tyr	Asp	Leu	Tyr	Leu	Pro
				5					10					15
Lys	Ser	Trp	Thr	Ile	Gln	Val	Gly	Leu	Val	Ser	Leu	Leu	Asp	Asn

				20					25					30
Pro	Ala	Pro	Ser	His	Leu	Val	Glu	Lys	Ile	Val	Tyr	His	Ser	Lys
				35					40					45
Tyr	Lys	Pro	Lys	Arg	Leu	Gly	Asn	Asp	Ile	Ala	Leu	Leu		
				50					55					

<210> 7
 <211> 328
 <212> DNA
 <213> Artificial sequence
 <220>
 <223> nucleotide sequence of the subclone containing
 the 300 bp band from the PCR product for
 TADG-12 variant, which contains an additional
 insert of 133 bases
 <400> 7

gggtggtgac	ggcggcgcac	tgtgtttatg	agattgtagc	tcctagagaa	50
agggcagaca	gaagaggaag	gaagctcctg	tgctggagga	aaccacaaa	100
aatgaaagga	cctagacctt	cccatagcta	attccagtgg	accatgttat	150
ggcagataca	ggcttgtacc	tccccaagtc	atggaccatc	caggtgggtc	200
tagtttcct	gttggaaca	ccagcccat	cccacttgg	ggagaagatt	250
gtctaccaca	gcaagtacaa	gccaaagagg	ctgggcaacg	acatcgccct	300
cctaatacact	agtgcggccg	cctgcagg			328

<210> 8
 <211> 42
 <212> PRT
 <213> Artificial sequence
 <220>
 <223> deduced amino acid sequence of the 300 bp band
 from the PCR product for TADG-12 variant, which is
 a truncated form of TADG-12
 <400> 8

Val	Val	Thr	Ala	Ala	His	Cys	Val	Tyr	Glu	Ile	Val	Ala	Pro	Arg
				5					10					15
Glu	Arg	Ala	Asp	Arg	Arg	Gly	Arg	Lys	Leu	Leu	Cys	Trp	Arg	Lys
				20					25					30
Pro	Thr	Lys	Met	Lys	Gly	Pro	Arg	Pro	Ser	His	Ser			
				35					40					

<210> 9
 <211> 34
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> LDLR-A domain of the complement subunit C8
 (Comp8)
 <400> 9

Cys	Glu	Gly	Phe	Val	Cys	Ala	Gln	Thr	Gly	Arg	Cys	Val	Asn	Arg
				5					10					15
Arg	Leu	Leu	Cys	Asn	Gly	Asp	Asn	Asp	Cys	Gly	Asp	Gln	Ser	Asp
				20					25					30

Glu Ala Asn Cys

<210> 10
 <211> 34
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> LDLR-A domain of the serine protease
 matriptase (Matr)
 <400> 10

Cys Pro Gly Gln Phe Thr Cys Arg Thr Gly Arg Cys Ile Arg Lys
 5 10 15
 Glu Leu Arg Cys Asp Gly Trp Ala Asp Cys Thr Asp His Ser Asp
 20 25 30
 Glu Leu Asn Cys

<210> 11
 <211> 37
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> LDLR-A domain of the glycoprotein GP300
 (Gp300-1)
 <400> 11

Cys Gln Gln Gly Tyr Phe Lys Cys Gln Ser Glu Gly Gln Cys Ile
 5 10 15
 Pro Ser Ser Trp Val Cys Asp Gln Asp Gln Asp Cys Asp Asp Gly
 20 25 30
 Ser Asp Glu Arg Gln Asp Cys
 35

<210> 12
 <211> 35
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> LDLR-A domain of the glycoprotein GP300
 (Gp300-2)
 <400> 12

Cys Ser Ser His Gln Ile Thr Cys Ser Asn Gly Gln Cys Ile Pro
 5 10 15
 Ser Glu Tyr Arg Cys Asp His Val Arg Asp Cys Pro Asp Gly Ala
 20 25 30
 Asp Glu Asn Asp Cys
 35

<210> 13
 <211> 35

<212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <222> 74...108
 <223> LDLR-A domain of TADG-12
 <400> 13

Cys	Ser	Gly	Lys	Tyr	Arg	Cys	Arg	Ser	Ser	Phe	Lys	Cys	Ile	Glu
				5					10					15
Leu	Ile	Thr	Arg	Cys	Asp	Gly	Val	Ser	Asp	Cys	Lys	Asp	Gly	Glu
				20					25					30
Asp	Glu	Tyr	Arg	Cys										
				35										

<210> 14
 <211> 36
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> LDLR-A domain of the serine protease TMPRSS2
 Tmprss2
 <400> 14

Cys	Ser	Asn	Ser	Gly	Ile	Glu	Cys	Asp	Ser	Ser	Gly	Thr	Cys	Ile
				5					10					15
Asn	Pro	Ser	Asn	Trp	Cys	Asp	Gly	Val	Ser	His	Cys	Pro	Gly	Gly
				20					25					30
Glu	Asp	Glu	Asn	Arg	Cys									
				35										

<210> 15
 <211> 101
 <212> PRT
 <213> *Bos taurus*
 <220>
 <221> DOMAIN
 <223> SRCR domain of bovine enterokinase (BovEntk)
 <400> 15

Val	Arg	Leu	Val	Gly	Gly	Ser	Gly	Pro	His	Glu	Gly	Arg	Val	Glu
				5					10					15
Ile	Phe	His	Glu	Gly	Gln	Trp	Gly	Thr	Val	Cys	Asp	Asp	Arg	Trp
				20					25					30
Glu	Leu	Arg	Gly	Gly	Leu	Val	Val	Cys	Arg	Ser	Leu	Gly	Tyr	Lys
				35					40					45
Gly	Val	Gln	Ser	Val	His	Lys	Arg	Ala	Tyr	Phe	Gly	Lys	Gly	Thr
				50					55					60
Gly	Pro	Ile	Trp	Leu	Asn	Glu	Val	Phe	Cys	Phe	Gly	Lys	Glu	Ser
				65					70					75
Ser	Ile	Glu	Glu	Cys	Arg	Ile	Arg	Gln	Trp	Gly	Val	Arg	Ala	Cys
				80					85					90
Ser	His	Asp	Glu	Asp	Ala	Gly	Val	Thr	Cys	Thr				
				95					100					

<210> 16
 <211> 101
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> SRCR domain of human macrophage scavenger
 receptor (MacSR)
 <400> 16

Val	Arg	Leu	Val	Gly	Gly	Ser	Gly	Pro	His	Glu	Gly	Arg	Val	Glu
				5					10					15
Ile	Leu	His	Ser	Gly	Gln	Trp	Gly	Thr	Ile	Cys	Asp	Asp	Arg	Trp
				20					25					30
Glu	Val	Arg	Val	Gly	Gln	Val	Val	Cys	Arg	Ser	Leu	Gly	Tyr	Pro
				35					40					45
Gly	Val	Gln	Ala	Val	His	Lys	Ala	Ala	His	Phe	Gly	Gln	Gly	Thr
				50					55					60
Gly	Pro	Ile	Trp	Leu	Asn	Glu	Val	Phe	Cys	Phe	Gly	Arg	Glu	Ser
				65					70					75
Ser	Ile	Glu	Glu	Cys	Lys	Ile	Arg	Gln	Trp	Gly	Thr	Arg	Ala	Cys
				80					85					90
Ser	His	Ser	Glu	Asp	Ala	Gly	Val	Thr	Cys	Thr				
				95					100					

<210> 17
 <211> 98
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <222> 109...206
 <223> SRCR domain of TADG-12 (TADG12)
 <400> 17

Val	Arg	Val	Gly	Gly	Gln	Asn	Ala	Val	Leu	Gln	Val	Phe	Thr	Ala
				5					10					15
Ala	Ser	Trp	Lys	Thr	Met	Cys	Ser	Asp	Asp	Trp	Lys	Gly	His	Tyr
				20					25					30
Ala	Asn	Val	Ala	Cys	Ala	Gln	Leu	Gly	Phe	Pro	Ser	Tyr	Val	Ser
				35					40					45
Ser	Asp	Asn	Leu	Arg	Val	Ser	Ser	Leu	Glu	Gly	Gln	Phe	Arg	Glu
				50					55					60
Glu	Phe	Val	Ser	Ile	Asp	His	Leu	Leu	Pro	Asp	Asp	Lys	Val	Thr
				65					70					75
Ala	Leu	His	His	Ser	Val	Tyr	Val	Arg	Glu	Gly	Cys	Ala	Ser	Gly
				80					85					90
His	Val	Val	Thr	Leu	Gln	Cys	Thr							
				95										

<210> 18
 <211> 94
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN

<223> SRCR domain of the serine protease TMPRSS2
(Tmprss2)
<400> 18

Val	Arg	Leu	Tyr	Gly	Pro	Asn	Phe	Ile	Leu	Gln	Met	Tyr	Ser	Ser	5	10	15
Gln	Arg	Lys	Ser	Trp	His	Pro	Val	Cys	Gln	Asp	Asp	Trp	Asn	Glu	20	25	30
Asn	Tyr	Gly	Arg	Ala	Ala	Cys	Arg	Asp	Met	Gly	Tyr	Lys	Asn	Asn	35	40	45
Phe	Tyr	Ser	Ser	Gln	Gly	Ile	Val	Asp	Asp	Ser	Gly	Ser	Thr	Ser	50	55	60
Phe	Met	Lys	Leu	Asn	Thr	Ser	Ala	Gly	Asn	Val	Asp	Ile	Tyr	Lys	65	70	75
Lys	Leu	Tyr	His	Ser	Asp	Ala	Cys	Ser	Ser	Lys	Ala	Val	Val	Ser	80	85	90
Leu	Arg	Cys	Leu														

<210> 19
<211> 90
<212> PRT
<213> *Homo sapiens*
<220>
<221> DOMAIN
<223> SRCR domain of human enterokinase (HumEntk)
<400> 19

Val	Arg	Phe	Phe	Asn	Gly	Thr	Thr	Asn	Asn	Asn	Gly	Leu	Val	Arg	5	10	15
Phe	Arg	Ile	Gln	Ser	Ile	Trp	His	Thr	Ala	Cys	Ala	Glu	Asn	Trp	20	25	30
Thr	Thr	Gln	Ile	Ser	Asn	Asp	Val	Cys	Gln	Leu	Leu	Gly	Leu	Gly	35	40	45
Ser	Gly	Asn	Ser	Ser	Lys	Pro	Ile	Phe	Ser	Thr	Asp	Gly	Gly	Pro	50	55	60
Phe	Val	Lys	Leu	Asn	Thr	Ala	Pro	Asp	Gly	His	Leu	Ile	Leu	Thr	65	70	75
Pro	Ser	Gln	Gln	Cys	Leu	Gln	Asp	Ser	Leu	Ile	Arg	Leu	Gln	Cys	80	85	90

<210> 20
<211> 149
<212> PRT
<213> *Homo sapiens*
<220>
<221> DOMAIN
<223> protease domain of protease M (ProM)
<400> 20

Leu	Trp	Val	Leu	Thr	Ala	Ala	His	Cys	Lys	Lys	Pro	Asn	Leu	Gln	5	10	15
Val	Phe	Leu	Gly	Lys	His	Asn	Leu	Arg	Gln	Arg	Glu	Ser	Ser	Gln	20	25	30
Glu	Gln	Ser	Ser	Val	Val	Arg	Ala	Val	Ile	His	Pro	Asp	Tyr	Asp	35	40	45
Ala	Ala	Ser	His	Asp	Gln	Asp	Ile	Met	Leu	Leu	Arg	Leu	Ala	Arg			

Pro	Ala	Lys	Leu	50	Ser	Glu	Leu	Ile	Gln	55	Pro	Leu	Pro	Leu	Glu	60
				65						70						75
Asp	Cys	Ser	Ala	80	Asn	Thr	Thr	Ser	Cys	85	His	Ile	Leu	Gly	Trp	90
Lys	Thr	Ala	Asp	95	Gly	Asp	Phe	Pro	Asp	100	Thr	Ile	Gln	Cys	Ala	105
Ile	His	Leu	Val	110	Ser	Arg	Glu	Glu	Cys	115	Glu	His	Ala	Tyr	Pro	120
Gln	Ile	Thr	Gln	125	Asn	Met	Leu	Cys	Ala	130	Gly	Asp	Glu	Lys	Tyr	135
Lys	Asp	Ser	Cys	140	Gln	Gly	Asp	Ser	Gly	145	Gly	Pro	Leu	Val	Cys	

<210> 21
 <211> 151
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> protease domain of trypsinogen I (Try1)
 <400> 21

Gln	Trp	Val	Val	5	Ser	Ala	Gly	His	Cys	10	Tyr	Lys	Ser	Arg	Ile	Gln
Val	Arg	Leu	Gly	20	Glu	His	Asn	Ile	Glu	25	Val	Leu	Glu	Gly	Asn	Glu
Gln	Phe	Ile	Asn	35	Ala	Ala	Lys	Ile	Ile	40	Arg	His	Pro	Gln	Tyr	45
Arg	Lys	Thr	Leu	50	Asn	Asn	Asp	Ile	Met	55	Leu	Ile	Lys	Leu	Ser	60
Arg	Ala	Val	Ile	65	Asn	Ala	Arg	Val	Ser	70	Thr	Ile	Ser	Leu	Pro	75
Ala	Pro	Pro	Ala	80	Thr	Gly	Thr	Lys	Cys	85	Leu	Ile	Ser	Gly	Trp	90
Asn	Thr	Ala	Ser	95	Ser	Gly	Ala	Asp	Tyr	100	Pro	Asp	Glu	Leu	Gln	105
Leu	Asp	Ala	Pro	110	Val	Leu	Ser	Gln	Ala	115	Lys	Cys	Glu	Ala	Ser	120
Pro	Gly	Lys	Ile	125	Thr	Ser	Asn	Met	Phe	130	Cys	Val	Gly	Phe	Leu	135
Gly	Gly	Lys	Asp	140	Ser	Cys	Gln	Gly	Asp	145	Ser	Gly	Gly	Pro	Val	150

Cys

<210> 22
 <211> 158
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> protease domain of plasma kallikrein (Kal)
 <400> 22

Gln Trp Val Leu Thr Ala Ala His Cys Phe Asp Gly Leu Pro Leu

Gln Asp Val Trp	Arg Ile Tyr Ser Gly	Ile Leu Asn Leu Ser Asp	5 10 15
Ile Thr Lys Asp	Thr Pro Phe Ser Gln	Ile Lys Glu Ile Ile Ile	20 25 30
His Gln Asn Tyr	Lys Val Ser Glu Gly	Asn His Asp Ile Ala Leu	35 40 45
Ile Lys Leu Gln	Ala Pro Leu Asn Tyr	Thr Glu Phe Gln Lys Pro	50 55 60
Ile Cys Leu Pro	Ser Lys Gly Asp Thr	Ser Thr Ile Tyr Thr Asn	65 70 75
Cys Trp Val Thr	Gly Trp Gly Phe Ser	Lys Glu Lys Gly Glu Ile	80 85 90
Gln Asn Ile Leu	Gln Lys Val Asn Ile	Pro Leu Val Thr Asn Glu	95 100 105
Glu Cys Gln Lys	Arg Tyr Gln Asp Tyr	Lys Ile Thr Gln Arg Met	110 115 120
Val Cys Ala Gly	Tyr Lys Glu Gly Gly	Lys Asp Ala Cys Lys Gly	125 130 135
Asp Ser Gly Gly	Pro Leu Val Cys		140 145 150
			155

<210> 23
 <211> 157
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> protease domain of TADG-12 (TADG12)
 <400> 23

Leu Trp Ile Ile	Thr Ala Ala His Cys Val Tyr Asp Leu Tyr Leu	5 10 15
Pro Lys Ser Trp	Thr Ile Gln Val Gly Leu Val Ser Leu Leu Asp	20 25 30
Asn Pro Ala Pro	Ser His Leu Val Glu Lys Ile Val Tyr His Ser	35 40 45
Lys Tyr Lys Pro	Lys Arg Leu Gly Asn Asp Ile Ala Leu Met Lys	50 55 60
Leu Ala Gly Pro	Leu Thr Phe Asn Glu Met Ile Gln Pro Val Cys	65 70 75
Leu Pro Asn Ser	Glu Glu Asn Phe Pro Asp Gly Lys Val Cys Trp	80 85 90
Thr Ser Gly Trp	Gly Ala Thr Glu Asp Gly Gly Asp Ala Ser Pro	95 100 105
Val Leu Asn His	Ala Ala Val Pro Leu Ile Ser Asn Lys Ile Cys	110 115 120
Asn His Arg Asp	Val Tyr Gly Gly Ile Ile Ser Pro Ser Met Leu	125 130 135
Cys Ala Gly Tyr	Leu Thr Gly Gly Val Asp Ser Cys Gln Gly Asp	140 145 150
Ser Gly Gly Pro	Leu Val Cys	155

<210> 24
 <211> 159

<212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> protease domain of TMPRSS2 (Tmprss2)
 <400> 24

Glu	Trp	Ile	Val	Thr	Ala	Ala	His	Cys	Val	Glu	Lys	Pro	Leu	Asn
				5					10					15
Asn	Pro	Trp	His	Trp	Thr	Ala	Phe	Ala	Gly	Ile	Leu	Arg	Gln	Ser
				20					25					30
Phe	Met	Phe	Tyr	Gly	Ala	Gly	Tyr	Gln	Val	Gln	Lys	Val	Ile	Ser
				35					40					45
His	Pro	Asn	Tyr	Asp	Ser	Lys	Thr	Lys	Asn	Asn	Asp	Ile	Ala	Leu
				50					55					60
Met	Lys	Leu	Gln	Lys	Pro	Leu	Thr	Phe	Asn	Asp	Leu	Val	Lys	Pro
				65					70					75
Val	Cys	Leu	Pro	Asn	Pro	Gly	Met	Met	Leu	Gln	Pro	Glu	Gln	Leu
				80					85					90
Cys	Trp	Ile	Ser	Gly	Trp	Gly	Ala	Thr	Glu	Glu	Lys	Gly	Lys	Thr
				95					100					105
Ser	Glu	Val	Leu	Asn	Ala	Ala	Lys	Val	Leu	Leu	Ile	Glu	Thr	Gln
				110					115					120
Arg	Cys	Asn	Ser	Arg	Tyr	Val	Tyr	Asp	Asn	Leu	Ile	Thr	Pro	Ala
				125					130					135
Met	Ile	Cys	Ala	Gly	Phe	Leu	Gln	Gly	Asn	Val	Asp	Ser	Cys	Gln
				140					145					150
Gly	Asp	Ser	Gly	Gly	Pro	Leu	Val	Thr						
				155										

<210> 25
 <211> 164
 <212> PRT
 <213> *Homo sapiens*
 <220>
 <221> DOMAIN
 <223> protease domain of Hepsin (Heps)
 <400> 25

Asp	Trp	Val	Leu	Thr	Ala	Ala	His	Cys	Phe	Pro	Glu	Arg	Asn	Arg
				5					10					15
Val	Leu	Ser	Arg	Trp	Arg	Val	Phe	Ala	Gly	Ala	Val	Ala	Gln	Ala
				20					25					30
Ser	Pro	His	Gly	Leu	Gln	Leu	Gly	Val	Gln	Ala	Val	Val	Tyr	His
				35					40					45
Gly	Gly	Tyr	Leu	Pro	Phe	Arg	Asp	Pro	Asn	Ser	Glu	Glu	Asn	Ser
				50					55					60
Asn	Asp	Ile	Ala	Leu	Val	His	Leu	Ser	Ser	Pro	Leu	Pro	Leu	Thr
				65					70					75
Glu	Tyr	Ile	Gln	Pro	Val	Cys	Leu	Pro	Ala	Ala	Gly	Gln	Ala	Leu
				80					85					90
Val	Asp	Gly	Lys	Ile	Cys	Thr	Val	Thr	Gly	Trp	Gly	Asn	Thr	Gln
				95					100					105
Tyr	Tyr	Gly	Gln	Gln	Ala	Gly	Val	Leu	Gln	Glu	Ala	Arg	Val	Pro
				110					115					120
Ile	Ile	Ser	Asn	Asp	Val	Cys	Asn	Gly	Ala	Asp	Phe	Tyr	Gly	Asn

	125		130		135
Gln Ile Lys Pro	Lys Met Phe Cys Ala	Gly Tyr Pro Glu Gly	Gly		
	140		145		150
Ile Asp Ala Cys	Gln Gly Asp Ser Gly	Gly Pro Phe Val Cys			
	155		160		

<210> 26
 <211> 23
 <212> DNA
 <213> Artificial sequence
 <220>
 <221> primer_bind
 <222> 6, 9, 12, 15, 18
 <223> forward redundant primer for the consensus
 sequences of amino acids surrounding the catalytic
 triad for serine proteases, n = inosine
 <400> 26

tgggtngtna cngcngcnca ytg 23

<210> 27
 <211> 20
 <212> DNA
 <213> Artificial sequence
 <220>
 <221> primer_bind
 <222> 3, 6, 9, 12, 15, 18
 <223> reverse redundant primer for the consensus
 sequences of amino acids surrounding the catalytic
 triad for serine proteases, n = inosine
 <400> 27

arnarngcna tntcnttncc 20

<210> 28
 <211> 20
 <212> DNA
 <213> Artificial sequence
 <220>
 <221> primer_bind
 <223> forward oligonucleotide primer for TADG-12
 used for quantitative PCR
 <400> 28

gaaacatgtc cttgctctcg 20

<210> 29
 <211> 20
 <212> DNA
 <213> Artificial sequence
 <220>
 <221> primer_bind
 <223> reverse oligonucleotide primer for TADG-12
 used for quantitative PCR
 <400> 29

actaacttcc acagcctcct 20

<210> 30
<211> 20
<212> DNA
<213> Artificial sequence
<220>
<221> primer_bind
<223> forward oligonucleotide primer for TADG-12
variant (TADG-12V) used for quantitative PCR
<400> 30

tccaggtggg tctagtttcc 20

<210> 31
<211> 20
<212> DNA
<213> Artificial sequence
<220>
<221> primer_bind
<223> reverse oligonucleotide primer for TADG-12
variant (TADG-12V) used for quantitative PCR
<400> 31

ctcttttggt tgtacttgct 20

<210> 32
<211> 20
<212> DNA
<213> Artificial sequence
<220>
<221> primer_bind
<223> forward oligonucleotide primer for β -tubulin
used as an internal control for quantitative PCR
<400> 32

cgcatacaacg tgtactacaa 20

<210> 33
<211> 20
<212> DNA
<213> Artificial sequence
<220>
<221> primer_bind
<223> reverse oligonucleotide primer for β -tubulin
used as an internal control for quantitative PCR
<400> 33

tacgagctgg tggactgaga 20

<210> 34
<211> 12
<212> PRT
<213> Artificial sequence
<220>
<223> a poly-lysine linked multiple antigen peptide

derived from the TADG-12 carboxy-terminal protein sequence, present in full length TADG-12, but not in TADG-12V

<400> 34

Trp Ile His Glu Gln Met Glu Arg Asp Leu Lys Thr
5 10

<210> 35
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 40...48
<223> TADG-12 peptide
<400> 35

Ile Leu Ser Leu Leu Pro Phe Glu Val
5

<210> 36
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 144...152
<223> TADG-12 peptide
<400> 36

Ala Gln Leu Gly Phe Pro Ser Tyr Val
5

<210> 37
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 225...233
<223> TADG-12 peptide
<400> 37

Leu Leu Ser Gln Trp Pro Trp Gln Ala
5

<210> 38
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 252...260
<223> TADG-12 peptide
<400> 38

Trp Ile Ile Thr Ala Ala His Cys Val
5

<210> 39
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 356...364
<223> TADG-12 peptide
<400> 39

Val Leu Asn His Ala Ala Val Pro Leu
5

<210> 40
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 176...184
<223> TADG-12 peptide
<400> 40

Leu Leu Pro Asp Asp Lys Val Thr Ala
5

<210> 41
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 13...21
<223> TADG-12 peptide
<400> 41

Phe Ser Phe Arg Ser Leu Phe Gly Leu
5

<210> 42
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 151...159
<223> TADG-12 peptide
<400> 42

Tyr Val Ser Ser Asp Asn Leu Arg Val
5

<210> 43
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 436...444
<223> TADG-12 peptide
<400> 43

Arg Val Thr Ser Phe Leu Asp Trp Ile
5

<210> 44
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 234...242
<223> TADG-12 peptide
<400> 44

Ser Leu Gln Phe Gln Gly Tyr His Leu
5

<210> 45
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 181...189
<223> TADG-12 peptide
<400> 45

Lys Val Thr Ala Leu His His Ser Val
5

<210> 46
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 183...191
<223> TADG-12 peptide
<400> 46

Thr Ala Leu His His Ser Val Tyr Val
5

<210> 47
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 411...419
<223> TADG-12 peptide
<400> 47

Arg Leu Trp Lys Leu Val Gly Ala Thr
5

<210> 48
<211> 9
<212> PRT
<213> *Homo sapiens*

<220>
<222> 60...68
<223> TADG-12 peptide
<400> 48

Leu Ile Leu Ala Leu Ala Ile Gly Leu
5

<210> 49
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 227...235
<223> TADG-12 peptide
<400> 49

Ser Gln Trp Pro Trp Gln Ala Ser Leu
5

<210> 50
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 301...309
<223> TADG-12 peptide
<400> 50

Arg Leu Gly Asn Asp Ile Ala Leu Met
5

<210> 51
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 307...315
<223> TADG-12 peptide
<400> 51

Ala Leu Met Lys Leu Ala Gly Pro Leu
5

<210> 52
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 262...270
<223> TADG-12 peptide
<400> 52

Asp Leu Tyr Leu Pro Lys Ser Trp Thr
5

<210> 53
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 416...424
<223> TADG-12 peptide
<400> 53

Leu Val Gly Ala Thr Ser Phe Gly Ile
5

<210> 54
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 54...62
<223> TADG-12 peptide
<400> 54

Ser Leu Gly Ile Ile Ala Leu Ile Leu
5

<210> 55
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 218...226
<223> TADG-12 peptide
<400> 55

Ile Val Gly Gly Asn Met Ser Leu Leu
5

<210> 56
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 35...43
<223> TADG-12 peptide
<400> 56

Ala Val Ala Ala Gln Ile Leu Ser Leu
5

<210> 57
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 271...279
<223> TADG-12 peptide
<400> 57

Ile Gln Val Gly Leu Val Ser Leu Leu
5

<210> 58
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 397...405
<223> TADG-12 peptide
<400> 58

Cys Gln Gly Asp Ser Gly Gly Pro Leu
5

<210> 59
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 270...278
<223> TADG-12 peptide
<400> 59

Thr Ile Gln Val Gly Leu Val Ser Leu
5

<210> 60
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 56...64
<223> TADG-12 peptide
<400> 60

Gly Ile Ile Ala Leu Ile Leu Ala Leu
5

<210> 61
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 110...118
<223> TADG-12 peptide
<400> 61

Arg Val Gly Gly Gln Asn Ala Val Leu
5

<210> 62
<211> 9
<212> PRT
<213> *Homo sapiens*

<220>
<222> 217...225
<223> TADG-12 peptide
<400> 62

Arg Ile Val Gly Gly Asn Met Ser Leu
5

<210> 63
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 130...138
<223> TADG-12 peptide
<400> 63

Cys Ser Asp Asp Trp Lys Gly His Tyr
5

<210> 64
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 8...16
<223> TADG-12 peptide
<400> 64

Ala Val Glu Ala Pro Phe Ser Phe Arg
5

<210> 65
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 328...336
<223> TADG-12 peptide
<400> 65

Asn Ser Glu Glu Asn Phe Pro Asp Gly
5

<210> 66
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 3...11
<223> TADG-12 peptide
<400> 66

Glu Asn Asp Pro Pro Ala Val Glu Ala
5

<210> 67
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 98...106
<223> TADG-12 peptide
<400> 67

Asp Cys Lys Asp Gly Glu Asp Glu Tyr
5

<210> 68
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 346...354
<223> TADG-12 peptide
<400> 68

Ala Thr Glu Asp Gly Gly Asp Ala Ser
5

<210> 69
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 360...368
<223> TADG-12 peptide
<400> 69

Ala Ala Val Pro Leu Ile Ser Asn Lys
5

<210> 70
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 153...161
<223> TADG-12 peptide
<400> 70

Ser Ser Asp Asn Leu Arg Val Ser Ser
5

<210> 71
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 182...190
<223> TADG-12 peptide
<400> 71

Val Thr Ala Leu His His Ser Val Tyr
5

<210> 72
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 143...151
<223> TADG-12 peptide
<400> 72

Cys Ala Gln Leu Gly Phe Pro Ser Tyr
5

<210> 73
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 259...267
<223> TADG-12 peptide
<400> 73

Cys Val Tyr Asp Leu Tyr Leu Pro Lys
5

<210> 74
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 369...377
<223> TADG-12 peptide
<400> 74

Ile Cys Asn His Arg Asp Val Tyr Gly
5

<210> 75
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 278...286
<223> TADG-12 peptide
<400> 75

Leu Leu Asp Asn Pro Ala Pro Ser His
5

<210> 76
<211> 9
<212> PRT
<213> *Homo sapiens*

<220>
<222> 426...434
<223> TADG-12 peptide
<400> 76

Cys Ala Glu Val Asn Lys Pro Gly Val
5

<210> 77
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 32...40
<223> TADG-12 peptide
<400> 77

Asp Ala Asp Ala Val Ala Ala Gln Ile
5

<210> 78
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 406...414
<223> TADG-12 peptide
<400> 78

Val Cys Gln Glu Arg Arg Leu Trp Lys
5

<210> 79
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 329...337
<223> TADG-12 peptide
<400> 79

Ser Glu Glu Asn Phe Pro Asp Gly Lys
5

<210> 80
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 303...311
<223> TADG-12 peptide
<400> 80

Gly Asn Asp Ile Ala Leu Met Lys Leu
5

<210> 81
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 127...135
<223> TADG-12 peptide
<400> 81

Lys Thr Met Cys Ser Asp Asp Trp Lys
5

<210> 82
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 440...448
<223> TADG-12 peptide
<400> 82

Phe Leu Asp Trp Ile His Glu Gln Met
5

<210> 83
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 433...441
<223> TADG-12 peptide
<400> 83

Val Tyr Thr Arg Val Thr Ser Phe Leu
5

<210> 84
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 263...271
<223> TADG-12 peptide
<400> 84

Leu Tyr Leu Pro Lys Ser Trp Thr Ile
5

<210> 85
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 169...177
<223> TADG-12 peptide
<400> 85

Glu Phe Val Ser Ile Asp His Leu Leu
5

<210> 86
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 296...304
<223> TADG-12 peptide
<400> 86

Lys Tyr Lys Pro Lys Arg Leu Gly Asn
5

<210> 87
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 16...24
<223> TADG-12 peptide
<400> 87

Arg Ser Leu Phe Gly Leu Asp Asp Leu
5

<210> 88
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 267...275
<223> TADG-12 peptide
<400> 88

Lys Ser Trp Thr Ile Gln Val Gly Leu
5

<210> 89
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 81...89
<223> TADG-12 peptide
<400> 89

Arg Ser Ser Phe Lys Cys Ile Glu Leu
5

<210> 90
<211> 9
<212> PRT

<213> *Homo sapiens*
<220>
<222> 375...383
<223> TADG-12 peptide
<400> 90

Val Tyr Gly Gly Ile Ile Ser Pro Ser
5

<210> 91
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 110...118
<223> TADG-12 peptide
<400> 91

Arg Val Gly Gly Gln Asn Ala Val Leu
5

<210> 92
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 189...197
<223> TADG-12 peptide
<400> 92

Val Tyr Val Arg Glu Gly Cys Ala Ser
5

<210> 93
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 165...173
<223> TADG-12 peptide
<400> 93

Gln Phe Arg Glu Glu Phe Val Ser Ile
5

<210> 94
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 10...18
<223> TADG-12 peptide
<400> 94

Glu Ala Pro Phe Ser Phe Arg Ser Leu
5

<210> 95
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 407...415
<223> TADG-12 peptide
<400> 95

Cys Gln Glu Arg Arg Leu Trp Lys Leu
5

<210> 96
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 381...389
<223> TADG-12 peptide
<400> 96

Ser Pro Ser Met Leu Cys Ala Gly Tyr
5

<210> 97
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 375...383
<223> TADG-12 peptide
<400> 97

Val Tyr Gly Gly Ile Ile Ser Pro Ser
5

<210> 98
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 381...389
<223> TADG-12 peptide
<400> 98

Ser Pro Ser Met Leu Cys Ala Gly Tyr
5

<210> 99
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 362...370
<223> TADG-12 peptide

<400> 99

Val Pro Leu Ile Ser Asn Lys Ile Cys
5

<210> 100

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 373...381

<223> TADG-12 peptide

<400> 100

Arg Asp Val Tyr Gly Gly Ile Ile Ser
5

<210> 101

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 283...291

<223> TADG-12 peptide

<400> 101

Ala Pro Ser His Leu Val Glu Lys Ile
5

<210> 102

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 177...185

<223> TADG-12 peptide

<400> 102

Leu Pro Asp Asp Lys Val Thr Ala Leu
5

<210> 103

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 47...55

<223> TADG-12 peptide

<400> 103

Glu Val Phe Ser Gln Ser Ser Ser Leu
5

<210> 104

<211> 9

<212> PRT

<213> *Homo sapiens*
<220>
<222> 36...44
<223> TADG-12 peptide
<400> 104

Val Ala Ala Gln Ile Leu Ser Leu Leu
5

<210> 105
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 255...263
<223> TADG-12 peptide
<400> 105

Thr Ala Ala His Cys Val Tyr Asp Leu
5

<210> 106
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 138...146
<223> TADG-12 peptide
<400> 106

Tyr Ala Asn Val Ala Cys Ala Gln Leu
5

<210> 107
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 195...203
<223> TADG-12 peptide
<400> 107

Cys Ala Ser Gly His Val Val Thr Leu
5

<210> 108
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 215...223
<223> TADG-12 peptide
<400> 108

Ser Ser Arg Ile Val Gly Gly Asn Met
5

<210> 109
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 298...306
<223> TADG-12 peptide
<400> 109

Lys Pro Lys Arg Leu Gly Asn Asp Ile
5

<210> 110
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 313...321
<223> TADG-12 peptide
<400> 110

Gly Pro Leu Thr Phe Asn Glu Met Ile
5

<210> 111
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 108...116
<223> TADG-12 peptide
<400> 111

Cys Val Arg Val Gly Gly Gln Asn Ala
5

<210> 112
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 294...302
<223> TADG-12 peptide
<400> 112

His Ser Lys Tyr Lys Pro Lys Arg Leu
5

<210> 113
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 265...273
<223> TADG-12 peptide

<400> 113

Leu Pro Lys Ser Trp Thr Ile Gln Val
5

<210> 114

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 88...96

<223> TADG-12 peptide

<400> 114

Glu Leu Ile Thr Arg Cys Asp Gly Val
5

<210> 115

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 79...87

<223> TADG-12 peptide

<400> 115

Arg Cys Arg Ser Ser Phe Lys Cys Ile
5

<210> 116

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 255...263

<223> TADG-12 peptide

<400> 116

Thr Ala Ala His Cys Val Tyr Asp Leu
5

<210> 117

<211> 9

<212> PRT

<213> *Homo sapiens*

<220>

<222> 207...215

<223> TADG-12 peptide

<400> 117

Ala Cys Gly His Arg Arg Gly Tyr Ser
5

<210> 118

<211> 9

<212> PRT

<213> *Homo sapiens*
<220>
<222> 154...162
<223> TADG-12 peptide
<400> 118

Ser Asp Asn Leu Arg Val Ser Ser Leu
5

<210> 119
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 300...308
<223> TADG-12 peptide
<400> 119

Lys Arg Leu Gly Asn Asp Ile Ala Leu
5

<210> 120
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 435...443
<223> TADG-12 peptide
<400> 120

Thr Arg Val Thr Ser Phe Leu Asp Trp
5

<210> 121
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 376...384
<223> TADG-12 peptide
<400> 121

Tyr Gly Gly Ile Ile Ser Pro Ser Met
5

<210> 122
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 410...418
<223> TADG-12 peptide
<400> 122

Arg Arg Leu Trp Lys Leu Val Gly Ala
5

<210> 123
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 210...218
<223> TADG-12 peptide
<400> 123

His Arg Arg Gly Tyr Ser Ser Arg Ile
5

<210> 124
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 109...117
<223> TADG-12 peptide
<400> 124

Val Arg Val Gly Gly Gln Asn Ala Val
5

<210> 125
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 191...199
<223> TADG-12 peptide
<400> 125

Val Arg Glu Gly Cys Ala Ser Gly His
5

<210> 126
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 78...86
<223> TADG-12 peptide
<400> 126

Tyr Arg Cys Arg Ser Ser Phe Lys Cys
5

<210> 127
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 113...121
<223> TADG-12 peptide

<400> 127
Gly Gln Asn Ala Val Leu Gln Val Phe
5

<210> 128
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 91...99
<223> TADG-12 peptide
<400> 128

Thr Arg Cys Asp Gly Val Ser Asp Cys
5

<210> 129
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 38...46
<223> TADG-12 peptide
<400> 129

Ala Gln Ile Leu Ser Leu Leu Pro Phe
5

<210> 130
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 211...219
<223> TADG-12 peptide
<400> 130

Arg Arg Gly Tyr Ser Ser Arg Ile Val
5

<210> 131
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 216...224
<223> TADG-12 peptide
<400> 131

Ser Arg Ile Val Gly Gly Asn Met Ser
5

<210> 132
<211> 9
<212> PRT

<213> *Homo sapiens*
<220>
<222> 118...126
<223> TADG-12 peptide
<400> 132

Leu Gln Val Phe Thr Ala Ala Ser Trp
5

<210> 133
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 370...378
<223> TADG-12 peptide
<400> 133

Cys Asn His Arg Asp Val Tyr Gly Gly
5

<210> 134
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 393...401
<223> TADG-12 peptide
<400> 134

Gly Val Asp Ser Cys Gln Gly Asp Ser
5

<210> 135
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 235...243
<223> TADG-12 peptide
<400> 135

Leu Gln Phe Gln Gly Tyr His Leu Cys
5

<210> 136
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 427...435
<223> TADG-12 peptide
<400> 136

Ala Glu Val Asn Lys Pro Gly Val Tyr
5

<210> 137
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 162...170
<223> TADG-12 peptide
<400> 137

Leu Glu Gly Gln Phe Arg Glu Glu Phe
5

<210> 138
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 9...17
<223> TADG-12 peptide
<400> 138

Val Glu Ala Pro Phe Ser Phe Arg Ser
5

<210> 139
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 318...326
<223> TADG-12 peptide
<400> 139

Asn Glu Met Ile Gln Pro Val Cys Leu
5

<210> 140
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 256...264
<223> TADG-12 peptide
<400> 140

Ala Ala His Cys Val Tyr Asp Leu Tyr
5

<210> 141
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 46...54
<223> TADG-12 peptide

<400> 141
Phe Glu Val Phe Ser Gln Ser Ser Ser
5

<210> 142
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 64...72
<223> TADG-12 peptide
<400> 142

Leu Ala Ile Gly Leu Gly Ile His Phe
5

<210> 143
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 192...200
<223> TADG-12 peptide
<400> 143

Arg Glu Gly Cys Ala Ser Gly His Val
5

<210> 144
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 330...338
<223> TADG-12 peptide
<400> 144

Glu Glu Asn Phe Pro Asp Gly Lys Val
5

<210> 145
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 182...190
<223> TADG-12 peptide
<400> 145

Val Thr Ala Leu His His Ser Val Tyr
5

<210> 146
<211> 9
<212> PRT

<213> *Homo sapiens*
<220>
<222> 408...416
<223> TADG-12 peptide
<400> 146

Gln Glu Arg Arg Leu Trp Lys Leu Val
5

<210> 147
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 206...214
<223> TADG-12 peptide
<400> 147

Thr Ala Cys Gly His Arg Arg Gly Tyr
5

<210> 148
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 5...13
<223> TADG-12 peptide
<400> 148

Asp Pro Pro Ala Val Glu Ala Pro Phe
5

<210> 149
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 261...269
<223> TADG-12 peptide
<400> 149

Tyr Asp Leu Tyr Leu Pro Lys Ser Trp
5

<210> 150
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 33...41
<223> TADG-12 peptide
<400> 150

Ala Asp Ala Val Ala Ala Gln Ile Leu
5

<210> 151
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 168...176
<223> TADG-12 peptide
<400> 151

Glu Glu Phe Val Ser Ile Asp His Leu
5

<210> 152
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 304...312
<223> TADG-12 peptide
<400> 152

Asn Asp Ile Ala Leu Met Lys Leu Ala
5

<210> 153
<211> 9
<212> PRT
<213> *Homo sapiens*
<220>
<222> 104...112
<223> TADG-12 peptide
<400> 153

Asp Glu Tyr Arg Cys Val Arg Val Gly
5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/05612**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(7) :C07K 14/435, 14/705; A61K 38/03, 38/08, 38/17

US CL :530/350; 514/2

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 530/350; 514/2

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Medline, Biosis, WestElectronic data base consulted during the international search (name of data base and, where practicable, search terms used)
protein and nucleic acids databases**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TANIMOTO et al. Cloning and expression of TADG-15, a novel serine: protease expressed in ovarian cancer. Proceedings of the American Association for Cancer Research. March 1998, Vol. 39, page 648, abstract #4414, see entire abstract.	1, 12, 18-21 23
X	O'BRIEN et al. Cloning and expression TADG-15, a novel serine protease expressed in ovarian cancer. Tumor Biology. 1998, Supplement 2, page 33, abstract 0-42, see entire abstract.	1, 12, 18-21, 23
X	WO 98/41656 A1 (THE BOARD OF TRUSTEES OF THE UNIVERSITY OF ARKANSAS) 24 September 1998, claim 5, page 8.	22
X, P	Database Genecore version 4.5. Accession number AW104113, NCI-CGAP, 'National Cancer Institute, Cancer Genome Anatomy Project (CGAP), Tumor Gene Index,' sequence listing, 20 October 1999, see sequence listing.	1, 2

☐ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

* Special categories of cited documents:	*T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A document defining the general state of the art which is not considered to be of particular relevance	*X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E earlier document published on or after the international filing date	*Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*A* document member of the same patent family
O document referring to an oral disclosure, use, exhibition or other means	
P document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

04 JUNE 2000

Date of mailing of the international search report

06 JUL 2000

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

KAREN A. CANELLA

Telephone No. (703) 308-1235

Exhibit 26

Role of NH₂-terminal Positively Charged Residues in Establishing Membrane Protein Topology*

(Received for publication, April 14, 1993, and in revised form, May 21, 1993)

Griffith D. Parks† and Robert A. Lamb§

From the Howard Hughes Medical Institute and Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, Illinois 60208-3500

The paramyxovirus HN polypeptide is a model type II membrane protein, containing an internal uncleaved signal/anchor (S/A) and is oriented in the membrane with an NH₂-terminal cytoplasmic domain and COOH-terminal ectodomain (N_{ext} topology). To test the role of NH₂-terminal positively charged residues in directing the HN membrane topology, the 3 arginine (Arg) residues within the 17-amino-acid NH₂-terminal domain were systematically converted to a glutamine or glutamate, and the topology of the mutant proteins was examined after expression in CV-1 cells. The data indicate that: (i) each of the NH₂-terminal Arg residues contributes to the signal directing proper HN topology, since substitutions in any of the three positions resulted in ~13–23% inversion into the N_{ext} form; (ii) substitutions in the Arg directly flanking the signal/anchor domain resulted in slightly more inversion than those which were located more distally; and (iii) substitution with a negatively charged glutamate led to more inversion than did replacement with an uncharged glutamine. The effect of a single Arg to Glu substitution on the HN topology was enhanced when present in the context of a truncated NH₂-terminal cytoplasmic tail (3 residues). A comparison of the sequences flanking the signal/anchor of well documented type III proteins showed that the majority of these proteins contain a negatively charged residue flanking the NH₂-terminal side. An exception to this rule is the NB protein which contains a single positively charged Arg residue in this position. A chimeric protein containing the NB ectodomain and the HN S/A and HN ectodomain lead to a significant fraction (70%) of the chimeric protein adopting type II topology suggesting that the positive charge flanking the S/A domain is important for establishing type II topology. These data are discussed in the context of the loop model for the biogenesis of integral membrane proteins and the possible signals necessary for establishing differing orientations.

The ability of an integral membrane protein to function properly depends on the precise targeting of the cytoplasmic

and extracellular domains of the polypeptide to the correct side of the membrane. The signals directing a protein into a characteristic membrane topology are contained within the amino acid sequence of the polypeptide (Blobel, 1980) and must be very precise as it appears that all naturally occurring membrane proteins adopt only a single final orientation. The majority of known membrane proteins which span the lipid bilayer a single time are classified as type I proteins (nomenclature of von Heijne and Gavel, 1988), based on the presence of both an NH₂-terminal cleavable signal sequence which targets the nascent polypeptide to the ER¹ membrane through an interaction with the signal recognition particle (SRP; Walter and Lingappa, 1986) and a separate COOH-terminal hydrophobic domain which acts as a stop transfer domain (membrane anchor). These proteins have an extracellular NH₂-terminal domain and a cytoplasmic COOH-terminal tail (N_{ext} topology). A second class of membrane proteins has been found, with fewer known members than the type I membrane proteins, in which the proteins adopt the opposite orientation and have an NH₂-terminal cytoplasmic tail and a COOH-terminal ectodomain (N_{ext} topology). These type II proteins lack an NH₂-terminal cleavable signal sequence, but contain an internal hydrophobic signal/anchor (S/A) which serves a dual function: the signaling of the nascent polypeptide to the ER membrane and the subsequent anchoring of the polypeptide in the lipid bilayer. Examples of type II proteins include the transferrin receptor (Schneider *et al.*, 1984), asialoglycoprotein receptor (Spiess and Lodish, 1986), the family of Golgi-resident glycosyltransferases (Paulson and Colley, 1989), and the paramyxovirus HN protein (Hiebert *et al.*, 1985). The least common class of membrane proteins that span the lipid bilayer a single time are the type III proteins which also contain an internal uncleaved S/A, but these proteins have an extracellular NH₂-terminal domain and are in the N_{ext} orientation. Examples of type III proteins include the cytochrome P-450 proteins (Nelson and Strobel, 1988), the erythrocyte sialoglycoprotein β (High and Tanner, 1987), and the influenza A virus M₂ protein and influenza B virus NB protein (Lamb *et al.*, 1985; Williams and Lamb, 1986).

In contrast to the cleavable signal sequences of the type I membrane proteins which have been analyzed in detail both by amino acid comparison (von Heijne, 1984, 1985) and experimentally (*e.g.* Nothwehr and Gordon, 1989), relatively little is known about the structural features which distinguish the two types of membrane proteins with internal uncleaved S/A sequences. The type II and III proteins both appear to use the same SRP-mediated mechanism for targeting to the ER membrane (Lipp and Dobberstein, 1986b; Hull *et al.*, 1988). However, the signals which direct the steps following

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† Associate of the Howard Hughes Medical Institute. Present address: Dept. of Microbiology and Immunology, Bowman Gray School of Medicine of Wake Forest University, Winston-Salem, NC 27157-1064.

§ Investigator of the Howard Hughes Medical Institute. To whom correspondence should be addressed: Dept. of Biochemistry, Molecular Biology and Cell Biology, 2153 Sheridan Rd., Evanston, IL 60208-3500. Tel.: 708-491-5433; Fax: 708-491-2467.

¹ The abbreviations used are: ER, endoplasmic reticulum; SRP, signal recognition particle; N-glycanase; peptide: *n*-glycosidase F; PAGE, polyacrylamide gel electrophoresis.

A

Mutant		% N _{exo}
WT*	M V N [*] A T E D A P V R A T C R V L F R S/A	0
10*	————— E —————	23
12*	————— Q —————	18
14*	————— E —————	23
15*	————— Q —————	14
16*	————— E —————	13
17*	————— Q —————	13
18*	————— E — E —————	56
19*	————— Q — E —————	42
20*	————— E ————— E —————	48
21*	————— Q ————— E —————	39
22*	————— E — E —————	44
23*	————— Q — E —————	36
24*	————— E — E — E —————	80
25*	————— Q — E — E —————	76

B

WT* 10* 12* 14* 15* 16* 17* 18* 19* 20* 21* 22* 23* 24* 25*



FIG. 1. Structure and expression of HN* arginine substitution mutants. A, schematic diagram of Arg substitution mutants. The amino acid sequence of the NH₂-terminal domain of HN WT* is shown in the *one letter code* with the HN signal/anchor domain (S/A) depicted as a *hatched box*. A *solid horizontal line* denotes sequence identity to WT* with glutamate (E) or glutamine (Q) substitutions shown

this interaction of the S/A with SRP and lead to exclusively the N_{ext} or N_{int} topology have not been determined. Hydrophobicity appears to be the only structural requirement for an uncleaved S/A to function in the targeting and anchoring of a polypeptide (Audigier *et al.*, 1987; Parks *et al.*, 1989; Zerial *et al.*, 1987). As such, the analysis of topogenic sequences of type II and III proteins has focused on residues flanking the S/A domain, and it has been shown that these two types of proteins can be inverted in the membrane by complete exchange of NH_2 - or COOH -terminal S/A-flanking regions (Hauptle *et al.*, 1989; Parks *et al.*, 1989; Parks and Lamb, 1991). On the basis of a theoretical analysis, based on amino acid sequences available from databases and examining amino acid sequences flanking S/A domains, two different hypotheses have been proposed to explain the orientation of type II and III integral membrane proteins. (a) The "charge difference" rule (Hartmann *et al.*, 1989) proposed that when the differences in the sum of positive and negative charges within 15 residues of the NH_2 - and COOH -terminal sides of the S/A domain was calculated, the more positive side was cytoplasmic, in the manner of a dipole moment. (b) The "positive inside" rule (von Heijne, 1986; von Heijne and Gavel, 1988) proposed that the topology of the protein is governed by positive charges alone, and the domain containing the most positive charges is cytoplasmic. However, in the case of two different type II proteins, data obtained from a systematic mutational analysis did not support either the charge difference rule or the positive inside rule (Beltzer *et al.*, 1991; Parks and Lamb, 1991). The experimental data indicated that positive charges in the NH_2 -terminal domain of type II proteins play a pivotal role in directing the N_{ext} topology, since it has been shown that the removal of positive charges from the NH_2 -terminal S/A-flanking region leads to inversion of type II proteins into the N_{int} orientation, while the addition of positive charges to the COOH -terminal S/A-flanking region alone has little effect on topology (Beltzer *et al.*, 1991; Parks and Lamb, 1991).

In an analysis of charge-altered HN mutants (Parks and Lamb, 1991), it was proposed that the HN orientation signal is composed at least in part by a positively charged residue directly flanking the NH_2 -terminal side of the S/A. However, the potential role of positively charged residues located more distal to the S/A was not tested, and it has been postulated that these residues may also contribute to the orientation signal (High and Dobberstein, 1992). Here we report a systematic mutational analysis of the NH_2 -terminal positively charged residues of the HN protein cytoplasmic tail and their effect on HN orientation. The data indicate that each of the 3 NH_2 -terminal Arg residues contributes to the signal directing the type II topology, since charge-altering mutations in these residues lead to polypeptides which can adopt the inverted N_{int} orientation. The ability to invert the HN topology by these substitutions depends on the distance of the mutation from the S/A, as well as the charge of the substituting residue, and the effect of these alterations is enhanced when in the context of a truncated NH_2 -terminal domain. These results are discussed in a model for the topogenic signals of type I, II, and III proteins.

MATERIALS AND METHODS

Cells—Monolayer cultures of CV-1 cells were grown in Dulbecco's modified Eagle's medium containing 10% fetal calf serum as described (Lamb and Lai, 1982).

Plasmid Construction and Mutagenesis—To construct a pGEM3 plasmid containing a bacteriophage T₇ RNA polymerase transcription terminator (pGEM3-term), the appropriate 570-base pair fragment was excised from pGemex-2 (Promega, Madison, WI) by digestion with *NaeI* and *HindIII* and inserted into the *NaeI* and *HindIII* sites of pGEM3. A cDNA clone of the SV5 HN protein gene (Hiebert *et al.*, 1985) was modified previously to encode the addition of a consensus site for *N*-linked glycosylation (Asn-Ala-Thr) near the NH_2 terminus of the protein (HN*; Parks and Lamb, 1991), and a fragment from this clone (encoding residues 1–81) was used as a source of starting materials for oligonucleotide-directed mutagenesis after inserting into a bacteriophage M13 vector as described (Parks *et al.*, 1989). Likewise, a cDNA clone encoding a deletion of 14 of 17 NH_2 -terminal residues (HNG1; Parks and Lamb, 1990) was used as starting material for the construction of mutants MVE and MVQ. Following mutagenesis, DNA fragments were excised from the replicative form of M13 by digestion with *EcoRI* and *PstI* and linked to a DNA fragment encoding HN residues 82–565 in pGEM3-term (Arg substitution mutants) or pGem11 (MVR, MVE, and MVQ) such that mRNA sense transcripts could be produced using the T₇ RNA polymerase promoter. Nucleotide sequences were confirmed by dideoxynucleotide chain-terminating sequencing (Sanger *et al.*, 1977).

To construct the gene encoding the chimeric protein NBHH, a cDNA fragment encoding a portion of the influenza virus B/Lee/40 segment 6 gene (bases 1–58; Shaw *et al.*, 1982) was fused to HN using standard polymerase chain reaction protocols to create the precise junction of the NB NH_2 -terminal domain and the HN S/A domain (Arg/Thr). The construction of the gene encoding the M₂/HN chimeric protein MgHH has been described previously (Parks *et al.*, 1989).

Isotopic Labeling of Polypeptides, Immunoprecipitation, N-Glycanase Digestions, Protease Treatment of Microsomal Membranes, and Polyacrylamide Gel Electrophoresis—Proteins were expressed in CV-1 cells as described (Parks and Lamb, 1991) using a modified version of the vaccinia virus/T₇ RNA polymerase system of Fuerst *et al.* (1986). Vaccinia virus vTF7-3-infected cells were transfected with pGEM plasmid DNA encoding the HN mutants and radiolabeled from 3.5 to 4.5 h postinfection with 20–50 $\mu\text{Ci/ml}$ Tran^{35S}label (ICN Radiochemicals Inc., Irvine, CA) in Dulbecco's modified Eagle's medium lacking cysteine and methionine. Radiolabeled cells were washed in phosphate-buffered saline before lysis in 1% SDS. Immunoprecipitation of proteins from cell extracts with antisera to denatured HN (HN antisera) was as described previously (Erickson and Blobel, 1979; Ng *et al.*, 1990). Deglycosylation of proteins by treatment with peptide:N-glycosidase F (*N*-glycanase) was carried out as described (Williams and Lamb, 1986). Microsomal membranes were prepared from vaccinia virus-infected cells by Dounce homogenization (Adams and Rose, 1985) and analyzed by trypsin digestion as described previously (Parks *et al.*, 1989). Samples were analyzed by SDS-PAGE on 10% polyacrylamide gels, followed by fluorography (Lamb and Choppin, 1976). Autoradiograms were quantitated using a Molecular Dynamics model 400 series Phosphorimager (Sunnyvale, CA), and represent the average of at least two experiments.

Nomenclature—The nomenclature for type I–III proteins follows that of von Heijne and Gavel (1988). For the purposes of discussion, the borders of the S/A are operationally defined as the first charged residues located on either side of the first hydrophobic membrane-spanning region. The HN Arg substitution mutants (Fig. 1) are denoted by a numbering system which is a continuation of that used previously (Parks and Lamb, 1991). The HN cytoplasmic domain mutants MVR, MVE, and MVQ are named for the 3 residues which comprise the tail of these proteins. Hybrid proteins NBHH and MgHH are denoted by letters which represent the origin of the NH_2 -terminal domain (NB or M₂), with the transmembrane domain and cytoplasmic domain being derived from HN (H). The M₂ NH_2 -

below their position in the HN NH_2 -terminal domain. The location of the NH_2 -terminal consensus site for NH_2 -linked glycosylation is highlighted by an asterisk. Vertical arrows indicate the location of the altered Arg residues. Nomenclature for the mutants is described in the text. Percent N_{ext} values represent the average of at least two experiments. B, expression of Arg substitution mutants. CV-1 cells infected with vaccinia virus vTF7-3 were transfected with DNA plasmids encoding the Arg substitution mutants. Polypeptides were radiolabeled from 3.5–4.5 h postinfection with Tran^{35S}label, immunoprecipitated with HN antisera, and analyzed by SDS-PAGE. N_{ext} and N_{int} denote polypeptides with the WT HN and inverted membrane orientations, respectively.

terminal domain used (Mg) contains a site for addition of *N*-linked carbohydrate (Parks *et al.*, 1989).

RESULTS

Role of HN NH₂-terminal Arg Residues in Topogenesis—To examine experimentally the role of NH₂-terminal positively charged residues in the cytoplasmic tail of a type II integral membrane protein in directing membrane topology, a series of charge-altered mutants was produced in which the 3 NH₂-terminal Arg residues of HN were converted individually (Fig. 1A, mutants 10*, 12*, 14*-17*) or in combination (mutants 18*-25*) to a negatively charged glutamate (*E*) or uncharged glutamine (*Q*). As a means of monitoring directly expression in the N_{ext} form, each of these mutants also contained a single site for the addition of an *N*-linked carbohydrate residue which had been inserted near the HN NH₂ terminus (HN*, Parks and Lamb, 1991). It was anticipated that glycosylation of the NH₂-terminal domain of HN molecules inverted into the N_{ext} topology would result in a species with a slower electrophoretic mobility than that of unglycosylated HN and would allow for a distinction between molecules having the HN N_{ext} orientation (four accessible COOH-terminal glycosylation sites), *bona fide* inversion into the N_{ext} form (one accessible NH₂-terminal glycosylation site), and unglycosylated polypeptides which were defective in membrane targeting. The HN mutants were expressed to high levels by first infecting CV-1 cells with a recombinant vaccinia virus which synthesizes T₇ RNA polymerase (Fuerst *et al.*, 1986) and then transfecting the cells with DNA plasmids encoding the mutants under control of the T₇ promoter. After radiolabeling the cells with ³⁵S-labeled amino acids, polypeptides were immunoprecipitated from cell extracts using HN antisera and examined by SDS-PAGE.

As shown in Fig. 1B, each of the charge-altered mutants was synthesized to varying degrees as a mixture of two major polypeptides: a species with an electrophoretic mobility closely matching that of HN WT* (N_{ext}) and a faster migrating species denoted as N_{ext}. The slight differences in the electrophoretic mobilities of the mutant polypeptides most likely reflect aberrant migration due to their charge differences. With each mutant, a single species which migrated faster than the N_{ext} form was generated after removal of the carbohydrate residues by *N*-glycanase treatment, and this indicates that the two electrophoretic species observed in Fig. 1B are a single polypeptide chain backbone that differs by glycosylation (data not shown, but see Parks and Lamb, 1991). Trace amounts of polypeptides which migrate faster than the N_{ext} form are degradation products and have an electrophoretic mobility distinct from deglycosylated HN (data not shown). Pulse-labeling followed by chase experiments indicated that the N_{ext} and N_{ext} forms of mutant proteins were relatively stable (data not shown), and thus, a comparison of the fraction of each mutant found in the N_{ext} form is a valid measure of the relative effect of each mutation on topogenesis. Quantitation of several experiments by Phosphorimager analysis of the N_{ext} and N_{ext} species showed that 13–23% of each of the single Arg mutants was expressed in the inverted N_{ext} form (Fig. 1B, left panel).

When 2 of the 3 HN NH₂-terminal cytoplasmic domain Arg residues were mutated (Fig. 1B, middle panel, mutants 18*-23*), significantly more of the HN protein was inverted in the membrane in comparison to the single Arg substitutions. Within each pair of mutants, the substitution of an Arg residue by a negatively charged Glu resulted in slightly more efficient expression in the N_{ext} form than when the Arg was replaced by an uncharged Gln residue (*e.g.* compare mutant 18* with 19*). Furthermore, substitution of the Arg located

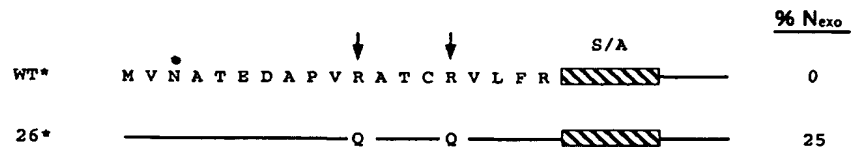
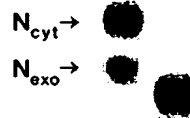
closest to the S/A led to greater expression in the N_{ext} form than did substitution of Arg residues which were more distal to the S/A, and this is most clearly seen by comparison of mutants 18* (56% N_{ext}) and 22* (44% N_{ext}). The largest inversion of the HN orientation was seen in the case of mutant 24* in which all of the Arg residues had been converted to Glu, and ~80% of this protein was oriented in the N_{ext} form (Fig. 1B, 24* lane). Taken together, these data suggest that substitution of each of the NH₂-terminal Arg residues leads to inversion of the HN type II topology, but that the positions closest to the S/A are more sensitive to these charge alterations.

To determine if a single Arg residue directly flanking the S/A was sufficient to direct the type II topology, a mutant HN* protein was constructed (Fig. 2, 26*) in which both Arg 11 and 15 were converted to uncharged Gln residues, leaving only Arg 19 which directly flanks the S/A. When the HN mutant 26* was expressed in CV-1 cells by the vaccinia virus T₇ RNA polymerase system described above, two major polypeptides were detected (Fig. 2, - lane), and both of these forms had an electrophoretic mobility which was slower than the single polypeptide produced after removal of the carbohydrate residues by treatment with *N*-glycanase (+ lane). Quantitation of the relative amounts of the two forms by Phosphorimager analysis showed that 25% of this protein was expressed in the N_{ext} orientation. Although the ability of each of the other 2 Arg residues to direct the N_{ext} orientation by themselves has not been tested, these data indicate that a single S/A-flanking positively charged residue is sufficient to direct 75% of the molecules into the type II topology. Furthermore, a comparison of the HN 26* mutant (25% N_{ext}) with the 22* mutant shown in Fig. 1B (44% N_{ext}) supports the above contention that the substitution of 2 Arg residues by a negatively charged Glu leads to greater inversion of HN than a substitution with uncharged Gln residues.

Effect of Arg Substitutions in the Context of a Truncated NH₂-terminal Domain—In the case of two other type II membrane proteins, IgCAT (Lipp and Dobberstein, 1986a) and the asialoglycoprotein receptor (Schmid and Spiess, 1988), truncations of the NH₂-terminal cytoplasmic tail result in molecules which were cleaved at a cryptic site in the S/A, and these processed polypeptides were soluble within the ER lumen. Analysis of the orientation of a cytoplasmic tail deletion mutant of a related HN protein (from Newcastle disease virus) suggested that the mutant protein was of mixed orientation (Wilson *et al.*, 1990). In contrast, when an SV5 HN mutant was constructed and expressed which has the NH₂-terminal domain truncated from 17 residues to the 3-residue tail MVR, a single major glycosylated species was detected (Fig. 3, MVR lanes). The available data indicate that the mutant MVR protein is integrated in the lipid bilayer (Parks and Lamb, 1990). We do not have a simple explanation for the difference in result obtained from two related HN cytoplasmic tail mutants except that the experiments differed in that *in vitro* and *in vivo* membrane integration was examined. As the data obtained with the MVR mutant were not complicated by a competing signal peptidase-like cleavage, it provided the opportunity to examine the effect of Arg substitutions within the context of the truncated MVR cytoplasmic tail.

Two mutants were constructed in which the single Arg residue in the MVR tail was converted to a Glu (*E*) or Gln (*Q*) residue to produce mutant proteins with NH₂-terminal domains of MVE and MVQ (Fig. 3). Expression of the MVQ mutant using the vaccinia virus system described above (MVQ lanes) produced a protein profile which matched that pro-

FIG. 2. Effect of a single NH₂-terminal S/A-flanking Arg residue on HN topology. CV-1 cells were infected with vaccinia vTTF-3 and transfected with a DNA plasmid encoding HN mutant 26*. After radiolabeling with Tran[³⁵S]label, polypeptides were immunoprecipitated from cell extracts with HN antisera. Immune complexes were divided into two portions, incubated with (+) or without (−) *N*-glycanase, and the polypeptides were examined by SDS-PAGE. The NH₂-terminal amino acid sequence of HN WT* is shown with the location of the 2 Arg residues converted to Gln to create the 26* mutant indicated by arrows.



duced by the MVR protein. For both MVR and MVQ, trace amounts of a faster migrating species were also observed (*lanes MVR-* and *MVQ-*), and these species have a different electrophoretic mobility than deglycosylated MVR and MVQ (+ *lanes*). It is thought likely that these species represent degradation products. In contrast, the MVE protein was synthesized as two major polypeptide species: one which migrated like the N_{cyt} form of MVR and a faster-migrating N_{exo} polypeptide with a mobility matching that of the single protein resulting from *N*-glycanase treatment (*MVE lanes*). Alkali treatment of microsomal membranes from cells expressing the MVE mutant did not remove either of these two protein species from the membrane (data not shown). However, the formal analysis of showing transmembrane topology by using proteases to trim a segment of the cytoplasmic tail could not be done because the small size of the cytoplasmic tail precludes a shift in electrophoretic mobility of the trimmed form on gels. Although these data do not provide formal proof that the NH_2 -terminal domain of the N_{exo} form of MVE has been fully translocated across the ER membrane, the strong association of both MVE species with the membrane suggests that the lack of glycosylation of the N_{exo} form was due to inversion into the type III orientation and was not due to defective integration into the membrane. Quantitation of the two forms of the MVE protein synthesized during a 1-h labeling period indicated that 50% of the MVE molecules adopted the inverted N_{exo} form. Mutant MVQ was not inverted in membranes as compared to when the same membrane-proximal mutation was made in the full 19-residue WT* tail (mutant 12*) (0 versus 18% in the N_{exo} form). A possible explanation is that the loss of the S/A-flanking positive charge in the MVQ mutant is compensated for by the positive charge contributed by the adjacent NH_2 terminus of this truncated

protein. As the MVE mutant contained the same membrane-proximal mutation as mutant 10* and yet led to different levels of protein-inversion (50 *versus* 23%), it lends further credence to the notion that other charge residues in the cytoplasmic domain are important in establishing orientation.

The NH₂-terminal Ectodomain of the Type III NB Protein Can Function as a Type II Cytoplasmic Tail—A compilation of the amino acid sequences of known type II membrane proteins shows that the vast majority of these proteins (~90%) have a residue with a positive charge (Arg or Lys) directly flanking the NH₂-terminal cytoplasmic side of the S/A (for compilations see reviews by Paulson and Colley, 1989; Hartmann *et al.*, 1989), and the importance of this positive charge for type II membrane protein topogenesis has been demonstrated experimentally (Parks and Lamb, 1991). For the small number of naturally existing proteins which are exceptions to this correlation and lack an NH₂-terminal positively charged S/A-flanking residue, it is possible that the presence of a negative charge in this position may be compensated for by a long stretch of positive charges located more distal (NH₂-terminal) to the S/A (*e.g.* neutral endopeptidase, Malfroy *et al.*, 1988); a suggestion made previously in formulating the positive inside rule for membrane protein topogenesis (von Heijne and Gavel, 1988) and supported by the experimental data shown in Fig. 1. In comparison to type II membrane proteins, there are relatively few known examples of the oppositely orientated type III proteins, but the vast majority have a negatively charged Glu or Asp residue directly flanking the NH₂-terminal side of the S/A (Fig. 4). One of the exceptions to this correlation is found with the influenza B virus NB protein (Williams and Lamb, 1986) which contains a single NH₂-terminal positively charged residue flanking the S/A domain. Earlier work has shown that when a chimeric

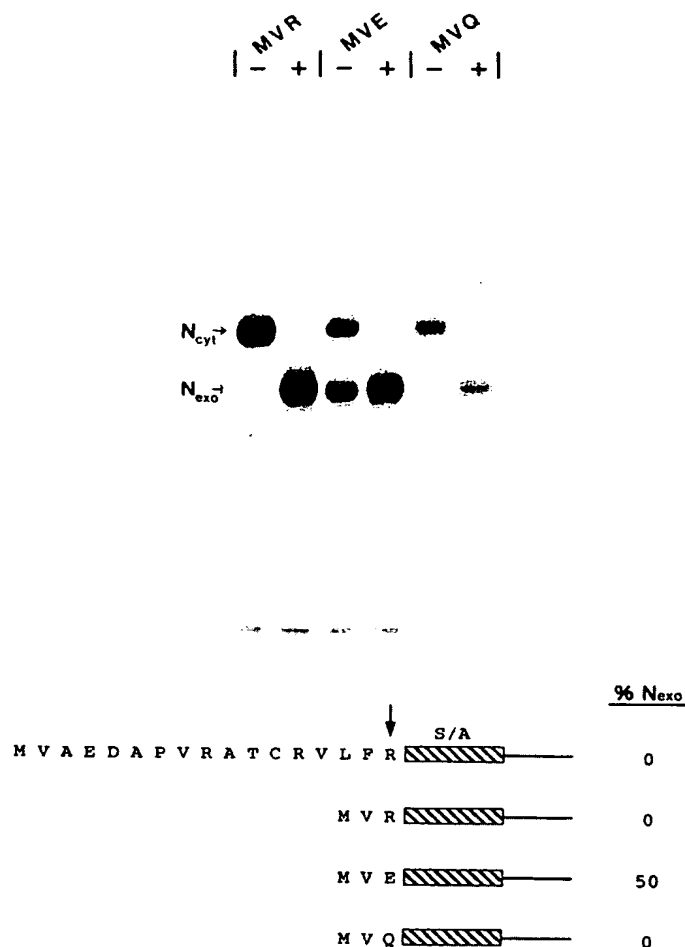


FIG. 3. The topological effect of charge alterations is enhanced in the context of a truncated HN NH₂-terminal domain. CV-1 cells infected with vaccinia virus vTF7-3 were transfected with plasmids encoding HN mutants MVR, MVE, or MVQ. Polypeptides were radiolabeled, immunoprecipitated with HN antisera, digested with (+) or without (-) N-glycanase, and analyzed by SDS-PAGE as described for Fig. 2. The NH₂-terminal sequence of the mutants is listed below that of HN, with the position of the altered Arg residue indicated by a vertical arrow.

protein MgHH, which was composed of the NH₂-terminal ectodomain of the type III M₂ protein linked to the HN S/A and COOH-terminal domains, was expressed the chimera integrated into membranes in two opposing orientations, but with the N_{exo} orientation predominating (Parks and Lamb, 1991 and see Fig. 5). As the NH₂-terminal domain of NB has a S/A domain-proximal positive charge but is functionally a type III ectodomain, it was of interest to determine which would be the predominating factor when this portion of the NB protein was linked to the HN S/A and COOH-terminal domains in a chimeric protein, NBHH.

The NBHH chimeric protein was expressed in CV-1 cells using the vaccinia T₇ system and was found as two predominant species (Fig. 5A, NBHH lanes): 70% as an N_{cyt} species with a mobility similar to that of the HN WT* protein (WT* lanes), and 30% as a faster migrating N_{exo} form. The difference in electrophoretic mobility between these two forms of NBHH was due to glycosylation (the N_{exo} form has two and the N_{cyt} form has four glycosylation sites) as only a single NBHH polypeptide species with identical mobility to deglycosylated WT* was detected after N-glycanase treatment (NBHH, + lanes). The membrane orientation of the two NBHH species

was further examined biochemically. Both NBHH N_{cyt} and N_{exo} forms were resistant to alkali extraction (data not shown), and the NBHH N_{cyt} form (like HN WT*) was protected from digestion by trypsin of microsomal membranes whereas the faster migrating NBHH N_{exo} form was susceptible to protease digestion (Fig. 5B). Taken together, these data suggest that the NB NH₂-terminal ectodomain is capable of acting as a cytoplasmic tail when linked to the HN S/A domain.

DISCUSSION

All nascent polypeptide chains use a common machinery for the targeting to the ER membrane (Walter and Lingappa, 1986), and yet by comparison very little amino acid identity is found among signal sequences. This is illustrated by a comparative sequence analysis (von Heijne, 1985) as well as experimentally, where it has been shown that seemingly random peptide sequences can function in targeting to the secretory pathway (Kaiser *et al.*, 1987; Paterson and Lamb, 1990). Likewise, the mechanism which follows this targeting to the membrane and leads to exclusively one orientation in the lipid bilayer must be precise and at the same time degenerate topogenic signals must be recognized, as there is little amino acid sequence identity among a variety of membrane proteins which have the same topology. Recent data indicate that charged residues are an important part of the signal for determining membrane protein topology (Beltzer *et al.*, 1991; Haeuptle *et al.*, 1989; Parks and Lamb, 1991).

The data obtained from a systematic analysis of the role of each of the HN NH₂-terminal Arg residues in determining the topology of the protein indicates that several conclusions can be drawn which address key features of membrane protein topology (reviewed in Boyd and Beckwith, 1990; High and Dobberstein, 1992) which although speculated on previously had not been examined by experiment. First, each of the 3 HN Arg residues contributes to the signal directing the N_{cyt} topology, with substitutions in the proximal S/A-flanking position leading to more inversion into the N_{exo} form than substitutions of the distal positions. It was shown previously that the S/A-flanking Arg residue is very important in establishing orientation. However, the charge alterations of this residue did not lead to complete inversion of HN in the membrane (Parks and Lamb, 1991). Thus, the observation that the inversion of HN was only partial can be explained by the presence of the other two NH₂-terminal Arg residues, and HN can be nearly completely inverted to the N_{exo} form (80%) by replacing all 3 Arg residues with Glu. The finding that the NB ectodomain can direct the N_{cyt} topology to approximately the same extent as the HN 26* mutant (which contains only a single S/A-flanking Arg) lends further support to the proposal that the exact sequence of a cytoplasmic tail is less critical for the generation of the type II topology than the position and number of positive charges (Parks and Lamb, 1991). Second, the relative importance of a given positively charged residue in contributing to the signal for topogenesis may depend on the length of the NH₂-terminal tail, since HN is inverted in the membrane to a greater extent when a charge alteration is introduced into a truncated tail than when it is introduced in the context of the full-length NH₂-terminal domain. Likewise, in the case of the asialoglycoprotein receptor (Beltzer *et al.*, 1991) 2 Arg to Asp substitutions lead to greater inversion in the membrane when introduced in the context of an NH₂-terminal tail which has been truncated from 40 (3% N_{exo}) to 11 residues (55% N_{exo}). Thus, the orientation signal may depend on the position and charge density of the positive charges, and these two factors could

PROTEIN	NH ₂	TM	COOH	REF.
R. cytochrome P450e		M E	R G H P K S R G N F P P	1
IBV 3C	M M N L L N K S L E E		R A L Q A F V Q A A D A	2
R. MinK	R R S Q L R D D S K L E		R S K K L E H S H D P F	3
IBV E1 protein	L D F E Q S V Q L F K E		R S K V I Y T L K M I V	4
M. LMu-CSF	P A P A L P L E D Q N E		R D T H R L T R T L N C	5
H. red/green opsin	Y T N S N S T R G P F E		K F K K L R H P L N W I	6
H. β -adrenergic rec.	A P D H D V T Q Q R D E		K F E R L Q T V T N Y F	7
H. β 1-adrenergic rec.	A S L L P P A S E S P E		K T P R L Q T L T N L F	8
B. opsin	S P F E A P Q Y Y L A E		H K K L R T P L N Y I L	9
Y. Sec63p	M P T N Y E Y D E A S E		E D G N S G K S K E F N	10
R. cytochrome P450 red.	V A E E V S L F S T T D		R K K K E E I P E F S K	11
H. glycophorin C	G R M E T S T P T I M D		R Y M Y R H K G T Y H T	12
B. substance K rec.	V M T D I N I S S G L D		H Q R M R T V T N Y F I	13
UR2 sarcoma virus ros	T P K T V D T V T S P D		H Q R W K S R K P A S T	14
rotavirus NS28	L M N S T L H T I L E D		H K A S I P T M K I A L	15
R. serotonin rec.	S S D G G R L F Q F P D		E K K L H N A T N Y F L	16
Influenza A virus M ₂	N E W G C R C N D S S D		D R L F F K C I Y R F F	17
H. blue opsin	P N Y H I A P R W V Y H		R Y K K L R Q P L N Y I	6
Influenza B virus NB	N C T N I N P I T H I R		K I F I N K N N C T N N	18
H. α ₂ -adrenergic rec.	W N G T E A P G G G A R		R A L K A P Q N L F L V	19
AEV v-erb-B	G P G L E G C P N G S K		R R R H I V R K R T L R	20

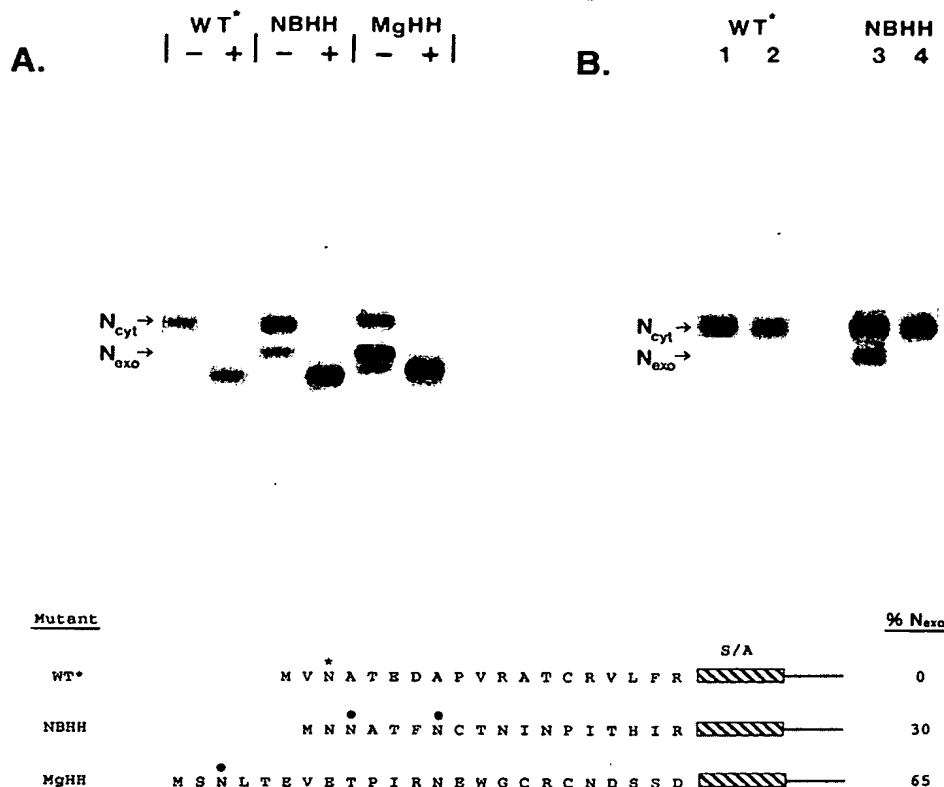
FIG. 4. Comparison of the amino acid sequence of type III proteins. The 12 amino acids flanking the amino- (NH₂) and carboxyl- (COOH) sides of the transmembrane domain (TM) of known type III (N_{ext}) proteins are listed in one letter code. The borders of the TM are operationally defined as the first charged residue on either side of the hydrophobic domain. In some instances (e.g. IBV E1 protein), the first transmembrane domain of a multispanning membrane protein has been shown to be an uncleaved S/A with the N_{ext} topology, and the relevant sequence of these proteins is included for completeness. This list may not be comprehensive, but includes those proteins for which there is reasonable biochemical evidence for type III topology. IBV, infectious bronchitis virus; LMu-CSF, long form of the multilineage colony-stimulating factor; rec., receptor; red., reductase; R., rat; M., murine; H., human; B., bovine; Y., yeast; AEV, avian erythroblastosis virus; UR2, avian sarcoma virus UR2. The references used are: 1) Nelson and Strobel, 1988; 2) Liu and Inglis, 1991; 3) Takumi *et al.*, 1988; 4) Machamer and Rose, 1987; 5) Haeuptle *et al.*, 1989; 6) Nathans *et al.*, 1986; 7) Schofield *et al.*, 1987; 8) Frielle *et al.*, 1987; 9) Nathans and Hogness, 1983; 10) Feldheim *et al.*, 1992; 11) Porter and Kasper, 1985; 12) High and Tanner, 1987; 13) Masu *et al.*, 1987; 14) Neckameyer *et al.*, 1985; 15) Bergmann *et al.*, 1989; 16) Julius *et al.*, 1988; 17) Lamb *et al.*, 1985; 18) Williams and Lamb, 1986; 19) Kobilka *et al.*, 1988; 20) Schatzman *et al.*, 1986.

explain those few examples of type II proteins which have a negatively charged residue flanking the NH₂-terminal side of the S/A (e.g. neutral endopeptidase, Malfroy *et al.*, 1988). Third, the substitution of Arg by a negatively charged Glu was a more potent inducer of inversion of HN orientation than was a replacement with an uncharged Gln (i.e. ~8–14% more in the N_{ext} form in the double Arg mutants). These data indicate that the inversion of HN orientation by these Arg substitutions was not due simply to lack of a positive charge and suggest that negative charges may act to promote translocation across the ER membrane. These observations are in contrast to the finding made for bacteria, where the orientation of an inner membrane protein can be reversed by the addition or removal of a single positively charged residue, but negative charges do not effect topology unless they are present in very high numbers (Nilsson and von Heijne, 1990; Andersson *et al.*, 1992).

A comparative analysis of the amino acids which comprise cleavable signal sequences indicates that these signals are

composed of three domains: a positively charged NH₂-terminal region, a central short stretch of hydrophobic residues, and a COOH-terminal region containing small polar residues which defines the site of cleavage by signal peptidase (von Heijne, 1984, 1985). The uncleaved S/A of a typical type II protein is structurally very similar to a type I signal sequence, and it has been shown experimentally that, except for the presence of a site for cleavage by signal peptidase in the type I proteins, these two signal sequences are functionally equivalent. It has been shown that a type II S/A can be converted to a cleavable signal sequence by NH₂-terminal alterations which expose a cryptic cleavage site (Lipp and Dobberstein, 1986a; Schmid and Spiess, 1988), and conversely it has been shown that a type I cleavable signal sequence can function as an uncleaved S/A when modified by extending the NH₂-terminal flanking domain and blocking the cleavage site (Shaw *et al.*, 1988). Based on these structural and functional similarities, it has been proposed that the type I and II proteins share a common mechanism for membrane integra-

FIG. 5. Expression and biochemical characterization of the NBHH hybrid protein. **A**, expression of NBHH. Vaccinia virus vTF7-3-infected cells were transfected with plasmid DNA encoding HN WT*, NBHH, or MgHH. Proteins were radiolabeled, immunoprecipitated with HN antisera, incubated with (+) or without (–) *N*-glycanase, and analyzed by SDS-PAGE as described in the legend to Fig. 2. The positions of the N_{cyt} and N_{exo} polypeptides are indicated. **B**, proteinase treatment of microsomal membranes from cells expressing WT* and NBHH. Vaccinia virus vTF7-3-infected cells were transfected with plasmids encoding HN WT* (lanes 1 and 2) or NBHH (lanes 3 and 4) and were radiolabeled with Tran[^{35}S]label. Crude microsomal membranes were prepared and treated with buffer (lanes 1 and 3) or with trypsin (lanes 2 and 4) as described previously (Parks *et al.*, 1989). Following centrifugation, samples were immunoprecipitated with HN antisera and analyzed by SDS-PAGE. The NH_2 -terminal sequence of HN WT* and of the chimeric NBHH and MgHH proteins is shown below, with a cross-hatched box and horizontal lines denoting the HN S/A and COOH-terminal ectodomain, respectively. The location of the consensus sites for *N*-linked glycosylation are highlighted by asterisks.



tion and topogenesis (von Heijne and Blomberg, 1979; Inouye and Halegona, 1980; Engelman and Steitz, 1981; Shaw *et al.*, 1988), with the nascent polypeptide being presented to the ER membrane as a loop structure formed by holding both NH_2 - and COOH -terminal sides of the signal sequence on the cytoplasmic side of the lipid bilayer with the NH_2 -terminal retention signal composed at least in part of positively charged residues (reviewed in High and Dobberstein, 1992).

In contrast to the establishment of type II protein orientation, the rules determining type III protein orientation remain enigmatic. Type III proteins depend on SRP for membrane targeting and integration (Hull *et al.*, 1988) and may be presented initially to the membrane as a loop structure (for a schematic diagram, see review by High and Dobberstein, 1992), but lacking the cytoplasmic retention signal the NH_2 terminus of these proteins would be translocated across the bilayer. As initially proposed to explain the topogenesis of the first N_{exo} transmembrane of opsin (Audigier *et al.*, 1987), the NH_2 -terminal region of all nascent membrane proteins (type I–III) may bind to an unrecognized factor to form the common loop structure, but for type III proteins this binding may be more readily dissociated leading to “flipping” of the NH_2 terminus across the ER membrane. The ability to vary the inversion of HN into the N_{exo} form by NH_2 -terminal charge alterations may reflect the degree of dissociation of the mutant NH_2 terminus from this putative binding factor, with positively charged residues being held more tightly than negatively charged residues. In the case of *Escherichia coli*, the acidic SecA protein appears to interact directly with positive charges in the signal sequence of nascent type I proteins during translocation across the cytoplasmic membrane (Akita *et al.*, 1990). Although a protein analogous to secA has not been identified to date in eukaryotic cells, recent cross-linking and reconstitution studies have led to the identification of several ER membrane proteins which may be directly involved

in forming an aqueous pore across membranes (reviewed in Rapoport, 1992). Thus, these proteins are candidates for interacting with the NH_2 -terminal positive charges of a nascent polypeptide chain. Alternatively, the type III proteins may employ a distinct topogenic mechanism, whereby the NH_2 terminus is not bound to form the transient loop structure but is presented to the ER membrane in a “head-on” configuration.

The experimental data described here indicate that it is possible to convert a type II protein into the N_{exo} topology by NH_2 -terminal charge alterations, and thus these data address indirectly the nature of the topogenic signals of naturally occurring type III proteins. Although experimentally a type III protein can be converted to a type II protein, by complete exchanges of S/A-flanking domains (Parks and Lamb, 1991), a direct systematic testing of the role of individual proximal and distal charges in generating the type III topology has yet to be performed. In the MgHH chimera, the type III Mg ectodomain which lacks a S/A-flanking-positively charged residue directed 65% of the molecule in the type III orientation, whereas in the NBHH chimera the type III NB ectodomain, which contains a positively charged residue flanking the S/A domain, directed 70% of the molecules in the opposing HN type II orientation. Thus, the signal for establishing type III topology may be complex and consist of the NH_2 -terminal ectodomain in conjunction with the S/A domain, and the artificial dividing of two parts of the signal in the chimera may explain the difference in the ability of the M₂ and NB type III ectodomains to function in directing the N_{exo} topology when linked to the HN S/A (MgHH and NBHH). This may also explain the observation that a chimeric protein can adopt dual orientations, a problem not found with naturally existing proteins. In the case of the type III cytochrome P-450 protein, it has been proposed that membrane topology is determined by a balance between the NH_2 -terminal charged residues and

the length of the hydrophobic signal (Sakaguchi *et al.*, 1992), with proteins in the N_{ext} topology requiring a longer hydrophobic stretch and fewer positive charges. Therefore, for type III proteins overlapping signals contributed by both the S/A and NH₂-terminal domains may act together to assure the precise steps in establishing membrane orientation.

Acknowledgments—We thank Margaret Shaughnessy for excellent technical assistance and Zhi-hai Ma and Oscar Valles for constructing some of the HN mutants as D99 projects at Northwestern University.

REFERENCES

- Adams, G. A., and Rose, J. K. (1985) *Mol. Cell Biol.* 5, 1442-1448
- Akita, M., Sasaki, S., Matsuyama, S.-I., and Mizushima, S. (1990) *J. Biol. Chem.* 265, 8164-8169
- Andersson, H., Bakker, E., and von Heijne, G. (1992) *J. Biol. Chem.* 267, 1491-1495
- Audigier, Y., Friedlander, M., and Blobel, G. (1987) *Proc. Natl. Acad. Sci. U. S. A.* 84, 5783-5787
- Beltzer, J. P., Fiedler, K., Fuhrer, C., Geffen, L., Handschin, C., Wessels, H. P., and Spiess, M. (1991) *J. Biol. Chem.* 266, 973-978
- Bergmann, C. C., Maass, D., Poruchynsky, M. S., Atkinson, P. H., and Bellamy, A. R. (1989) *EMBO J.* 8, 1543-1550
- Blobel, G. (1980) *Proc. Natl. Acad. Sci. U. S. A.* 77, 1496-1500
- Boyd, D., and Beckwith, J. (1990) *Cell* 62, 1031-1033
- Engelman, D. M., and Steitz, T. A. (1981) *Cell* 23, 411-422
- Erickson, A. H., and Blobel, G. (1979) *J. Biol. Chem.* 254, 11771-11774
- Feldheim, D., Rothblatt, J., and Schekman, R. (1992) *Mol. Cell Biol.* 12, 3288-3298
- Frielle, T., Collins, S., Daniel, K. W., Caron, M. G., Lefkowitz, R. J., and Kobilka, B. K. (1987) *Proc. Natl. Acad. Sci. U. S. A.* 84, 7920-7924
- Fuerst, T. R., Niles, E. G., Studier, F. W., and Moss, B. (1986) *Proc. Natl. Acad. Sci. U. S. A.* 83, 8122-8128
- Haupt, M., Flint, N., Gough, N. M., and Dobberstein, B. (1989) *J. Cell Biol.* 108, 1227-1236
- Hartmann, E., Rapoport, T. A., and Lodish, H. F. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 5786-5790
- Hiebert, S. W., Paterson, R. G., and Lamb, R. A. (1985) *J. Virol.* 54, 1-6
- High, S., and Dobberstein, B. (1992) *Curr. Opin. Cell Biol.* 4, 581-586
- High, S., and Tanner, M. J. A. (1987) *Biochem. J.* 243, 277-280
- Hull, J. D., Gilmore, R., and Lamb, R. A. (1988) *J. Cell Biol.* 106, 1489-1498
- Inouye, M., and Halagana, S. (1980) *CRC Crit. Rev. Biochem.* 7, 339-371
- Julius, D., MacDermott, A. B., Axel, R., Jessell, T. M. (1988) *Science* 241, 558-564
- Kaiser, C. A., Preuss, D., Grisafi, P., and Botstein, D. (1987) *Science* 235, 312-317
- Kobilka, B. K., Matsui, H., Kobilka, T. S., Yang-feng, T. L., Francke, U., Caron, M. G., Lefkowitz, R. J., and Regan, J. W. (1988) *Science* 238, 650-656
- Lamb, R. A., and Choppin, P. W. (1976) *Virology* 74, 504-519
- Lamb, R. A., and Lai, C.-J. (1982) *Virology* 123, 237-256
- Lamb, R. A., Zebedes, S. L., and Richardson, C. D. (1985) *Cell* 40, 627-633
- Lipp, J., and Dobberstein, B. (1986a) *Cell* 46, 1103-1112
- Lipp, J., and Dobberstein, B. (1986b) *J. Cell Biol.* 102, 2169-2175
- Liu, D. X., and Inglis, S. C. (1991) *Virology* 185, 911-917
- Machamer, C. E., and Rose, J. K. (1987) *J. Cell Biol.* 105, 1205-1214
- Malfroy, B., Kuang, W.-J., Seeburg, P. H., Mason, A. J., and Schofield, P. R. (1988) *FEBS Lett.* 229, 206-210
- Masu, Y., Nakayama, K., Tamaki, H., Harada, Y., Kuno, M., and Nakanishi, S. (1987) *Nature* 329, 838-838
- Nathans, J., and Hogness, D. S. (1983) *Cell* 34, 807-814
- Nathans, J., Thomas, D., and Hogness, D. S. (1986) *Science* 232, 193-202
- Neckameyer, W. S., and Wang, L.-H. (1985) *J. Virol.* 53, 878-884
- Nelson, D. R., and Strobel, H. W. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 6038-6050
- Nilsson, I., and von Heijne, G. (1990) *Cell* 62, 1135-1141
- Ng, D. T. W., Hiebert, S. W., and Lamb, R. A. (1990) *Mol. Cell Biol.* 10, 1989-2001
- Nothwehr, S. F., and Gordon, J. L. (1989) *J. Biol. Chem.* 264, 3979-3987
- Parks, G. D., Hull, J. D., and Lamb, R. A. (1989) *J. Cell Biol.* 109, 2023-2032
- Parks, G. D., and Lamb, R. A. (1990) *J. Virol.* 64, 3605-3616
- Parks, G. D., and Lamb, R. A. (1991) *Cell* 64, 777-787
- Paterson, R. G., and Lamb, R. A. (1990) *J. Cell Biol.* 110, 999-1011
- Porter, J. C., and Colley, K. J. (1989) *J. Biol. Chem.* 264, 17615-17618
- Porter, T. D., and Kasper, C. B. (1986) *Proc. Natl. Acad. Sci. U. S. A.* 83, 973-977
- Rapoport, T. A. (1992) *Science* 252, 931-936
- Sakaguchi, M., Tomiyoshi, R., Kuroiwa, T., Mihara, K., and Omura, T. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 16-19
- Sanger, F., Nicklin, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463-5467
- Schatzman, R. C., Evan, G. I., Privalaky, M. L., and Bishop, J. M. (1986) *Mol. Cell Biol.* 6, 1329-1333
- Schmid, S. R., and Spiess, M. (1988) *J. Biol. Chem.* 263, 16886-16891
- Schneider, C., Owen, M. J., Banville, D., and Williams, G. W. (1984) *Nature* 311, 675-678
- Schofield, P. R., Rhee, L. M., and Peralta, E. G. (1987) *Nucleic Acids Res.* 15, 3636
- Shaw, M. W., Lamb, R. A., Erickson, B. W., Briedis, D. J., and Choppin, P. W. (1982) *Proc. Natl. Acad. Sci. U. S. A.* 79, 6817-6821
- Shaw, A. S., Rottier, P. J. M., and Rose, J. K. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 7592-7596
- Spiess, M., and Lodish, H. F. (1986) *Cell* 44, 177-185
- Takumi, T., Ohkubo, H., and Nakanishi, S. (1988) *Science* 242, 1042-1045
- von Heijne, G. (1984) *EMBO J.* 3, 2315-2318
- von Heijne, G. (1985) *J. Mol. Biol.* 184, 99-105
- von Heijne, G. (1986) *EMBO J.* 5, 3021-3027
- von Heijne, G., and Blomberg, C. (1979) *Eur. J. Biochem.* 97, 175-181
- von Heijne, G., and Gavel, Y. (1988) *Eur. J. Biochem.* 174, 6711-6718
- Walker, P., and Lingappa, V. R. (1986) *Annu. Rev. Cell Biol.* 2, 499-516
- Williams, M. A., and Lamb, R. A. (1986) *Mol. Cell Biol.* 6, 4317-4328
- Wilson, C., Gilmore, R., and Morrison, T. (1990) *Mol. Cell Biol.* 10, 449-457
- Zerial, M., Huylebroeck, D., and Garoff, H. (1987) *Cell* 48, 147-155

Exhibit 27

Topology of Eukaryotic Type II Membrane Proteins: Importance of N-Terminal Positively Charged Residues Flanking the Hydrophobic Domain

Griffith D. Parks and Robert A. Lamb
Department of Biochemistry, Molecular Biology
and Cell Biology
and Howard Hughes Medical Institute
Northwestern University
Evanston, Illinois 60208-3500

Summary

We have tested the role of different charged residues flanking the sides of the signal/anchor (S/A) domain of a eukaryotic type II ($N_{\text{cyt}}C_{\text{exo}}$) integral membrane protein in determining its topology. The removal of positively charged residues on the N-terminal side of the S/A yields proteins with an inverted topology, while the addition of positively charged residues to only the C-terminal side has very little effect on orientation. Expression of chimeric proteins composed of domains from a type II protein (HN) and the oppositely oriented membrane protein M_2 indicates that the HN N-terminal domain is sufficient to confer a type II topology and that the M_2 N-terminal ectodomain can direct a type II topology when modified by adding positively charged residues. These data suggest that eukaryotic membrane protein topology is governed by the presence or absence of an N-terminal signal for retention in the cytoplasm that is composed in part of positive charges.

Introduction

The signals that direct membrane protein topology are precise, as it appears that almost all naturally occurring membrane proteins adopt only one final orientation, which is determined by the amino acid sequence of the polypeptide chain (Blobel, 1980). Integral membrane proteins that span the lipid bilayer a single time can be classified as type I, II, or III (nomenclature of von Heijne, 1988), and this is based on the nature of their hydrophobic domains and their orientation in membranes. Type I proteins contain an N-terminal cleavable signal sequence that targets the nascent polypeptide to the endoplasmic reticulum (ER) membrane (reviewed in Walter and Lingappa, 1986). The final $N_{\text{exo}}C_{\text{cyt}}$ topology of type I proteins is determined by cleavage in the ER lumen of the N-terminal signal sequence by signal peptidase (Evans et al., 1986), and their translocation across the membrane is halted by a C-terminal hydrophobic stop-transfer region that anchors the polypeptide in the lipid bilayer. Type I proteins constitute the major class of integral membrane proteins that span the membrane once. The type II proteins do not contain a cleavable signal sequence, but instead have a long stretch of hydrophobic residues, the signal/anchor domain (S/A), which serves the dual function of targeting and anchoring the polypeptide in the ER membrane with an $N_{\text{cyt}}C_{\text{exo}}$ topology. Examples of type II proteins include the transferrin receptor (Schneider et al., 1984), HLA-

associated invariant chain (Strubin et al., 1984), asialoglycoprotein receptor (Spiess and Lodish, 1985), and the paramyxovirus hemagglutinin-neuraminidase (HN) and SH proteins (Hiebert et al., 1985a, 1985b).

The type III proteins contain an internal uncleaved S/A but adopt the $N_{\text{exo}}C_{\text{cyt}}$ orientation; the known examples constitute a small group including gp74 v-erbB of avian erythroblastosis virus (Schatzman et al., 1986), erythrocyte sialoglycoprotein β (High and Tanner, 1987), cytochrome P450 (Sato et al., 1990), the influenza A virus M_2 protein, and the influenza B virus NB protein (Lamb et al., 1985; Williams and Lamb, 1988). Recent experimental evidence has provided support for the earlier speculation (von Heijne and Blomberg, 1979; Inouye and Halegoua, 1980; Engelman and Steitz, 1981) that the nascent polypeptide chain of type I and II proteins is inserted into the ER membrane by a common mechanism involving a hairpin loop structure, and that the final topology of these proteins is determined by the presence or absence, in type I and type II proteins, respectively, of a site in the N-terminal hydrophobic domain that can be cleaved by signal peptidase (Lipp and Dobberstein, 1986a; Shaw et al., 1988). Although the type III proteins, such as the influenza virus M_2 protein, appear to share the common SRP-mediated ER targeting mechanism found with type I and II proteins (Lipp and Dobberstein, 1986b; Hull et al., 1988), the detailed steps of their membrane insertion have not been characterized.

We are interested in determining the signals that direct the opposing membrane topologies of eukaryotic type II and type III integral membrane proteins and have used the HN and M_2 proteins as models. That the hydrophobic nature of the residues composing an S/A appear to be the only structural requirement for this domain to function in targeting and anchoring a polypeptide (Zerial et al., 1987) and that it has been shown that an S/A domain can be inverted in membranes without loss of function (Parks et al., 1989) suggest that sequences outside of the S/A of the type II and III proteins direct membrane orientation. Analysis of the sequences of known membrane proteins led to the proposal of the "positive inside rule" (von Heijne, 1986a; von Heijne and Gavel, 1988), in which membrane proteins orientate themselves with the most positively charged end in the cytoplasm. However, based on a recent comparison of the sequences of eukaryotic type II and III membrane proteins, a strong correlation between the sum of the charges flanking the S/A of a protein and its membrane topology has been identified (Hartmann et al., 1989). It was proposed that the net charge of the 15 residues flanking the two sides of the S/A directs the orientation of a nascent polypeptide and that the domain with the more positive overall charge is retained in the cytoplasm. Thus, this "charge difference" hypothesis predicts that it is not the absolute number of positive or negative charges flanking the S/A but the sum of the flanking charges that is important for directing the topology of the protein (Hartmann et al., 1989).

We report here experiments designed to examine the role of charged residues in determining topology. An HN cDNA clone was systematically altered by site-specific mutagenesis to introduce negatively charged residues into the N-terminal flanking region and positively charged residues into the C-terminal side. Analysis of the topology of the altered proteins expressed in CV-1 cells emphasizes the importance of N-terminal positive charges in the establishment of the HN topology. From analysis of the orientation of various chimeric molecules constructed from domains of HN and M₂ we suggest that the establishment of the type II N_{CYT}C_{EXO} topology is dependent on the presence of an N-terminal cytoplasmic retention signal, which is in part composed of positively charged residues, and that the opposing HN and M₂ orientations are governed by the presence or absence of this N-terminal signal in these two polypeptides.

Results

Construction of Charge-Altered HN Mutants

To determine if a charge difference between the N-terminal and C-terminal side of the S/A domain is a factor in establishing type II membrane topology, the cDNA clone of the model type II protein HN was systematically mutated by oligonucleotide-directed mutagenesis to generate a series of charge-altered HN proteins (Figure 1A). In this series of mutants, HN residues flanking both sides of the S/A domain were changed separately or in combination such that the sum of the charges within the N-terminal 15 residues was progressively more negative than that of the 15 C-terminal flanking residues. The charge difference rules (Hartmann et al., 1989) predict that each of these HN mutants should adopt an inverted N_{EXO}C_{CYT} topology and, because the only sites for N-linked glycosylation are in the C-terminal ectodomain (Hiebert et al., 1985a; Ng et al., 1990), these inverted molecules should be readily distinguishable from those proteins with the normal HN orientation by their lack of glycosylation.

Expression of Charge-Altered HN Proteins

To obtain a high level of expression of the mutant HN proteins, the vaccinia virus system of Fuerst et al. (1986) was employed. CV-1 cells infected with vaccinia virus vTF7-3, which expresses the bacteriophage T₇ RNA polymerase, were transfected with plasmid DNAs encoding the mutant proteins under control of the T₇ RNA polymerase promoter. After radiolabeling the cells for 1 hr with Tran[³⁵S] label, proteins were immunoprecipitated from cell extracts with HN antisera and examined by SDS-polyacrylamide gel electrophoresis (SDS-PAGE). Using this expression system, wild-type (WT) HN was synthesized as a single polypeptide of M_r ≈ 68,000 (Figure 1B, lane WT).

Expression of the HN mutants produced a protein profile that was significantly different from that of WT HN. The charge-altered mutants were synthesized to varying degrees as a mixture of two major polypeptides: a species with an electrophoretic mobility similar to that of WT HN, designated N_{CYT}, and a faster-migrating form (M_r ≈ 50,000, Figure 1B, lanes 1–9), designated N_{EXO}. Minor polypep-

tide species migrating faster than the N_{EXO} species are thought to be degradation products of WT HN as described previously (Ng et al., 1989). After treatment of the proteins with peptide:N-glycosidase F (N-glycanase), each of the mutants was detected as a single polypeptide with an electrophoretic mobility similar to that of the N_{EXO} protein (not shown), and this suggests that the N_{CYT} and N_{EXO} forms are a single polypeptide species that differ from each other by N-linked glycosylation. Further biochemical evidence that the N_{CYT} and N_{EXO} forms of altered HN molecules are integral membrane proteins with opposing orientations is presented below.











Pulse-labeling followed by chase protocols indicated that within a 1 hr period all the forms of the mutant HN were stable (data not shown), and thus quantitation of the amounts of the species that accumulate is a reasonable assay for determining the amounts in each orientation. Densitometric scanning of autoradiograms from several experiments indicated that the fraction of HN mutants 1 and 2 found in the N_{EXO} form was 12% and 30%, respectively (Figure 1A, % N_{EXO}), which suggests that the introduction of negatively charged residues to the N-terminal side of the S/A has an important effect on membrane orientation. In contrast, only 5%–6% of the total HN protein was synthesized as the N_{EXO} species in the case of mutants 3–5, which encode a normal N-terminal domain but are modified by the addition of positively charged residues to the C-terminal side of the S/A. Combinations of N- and C-terminal substitutions (mutants 6–9) had the largest effect on HN orientation, as an increasing fraction of the total HN protein was synthesized as the N_{EXO} species when N-terminally altered mutants 1 and 2 were further modified by the addition of positive charges to the C-terminal side of the S/A (Figure 1B, lanes 6–9). A minor species of unknown origin that migrates between the N_{CYT} and N_{EXO} forms was immunoprecipitated from cells expressing the most highly charge-altered proteins (lanes 7–9), but its presence does not affect the interpretation of the data. The inversion to the N_{EXO} form reached a maximum value of 75% with mutant 9, which encoded N- and C-terminal net charges of –2 and +4, respectively.

These data suggest that the normal HN orientation can be disrupted by alterations in charged residues flanking the S/A domain, and proteins can be produced that adopt more than one orientation. However, our data do not fulfill the predictions of the charge difference rules (Hartmann et al., 1989), as only proteins containing mutations on the N-terminal side of the S/A (mutants 1, 2, and 6–9) were significantly inverted in the membrane and the topology of the mutants altered only on the C-terminal side of the S/A (mutants 3–5) remained largely unaltered.

Biochemical Evidence for the Orientation of Charge-Altered HN Proteins

It was inferred from the electrophoretic mobility of the N_{EXO} protein that the C-terminal domain of these molecules, which contains the sites for N-linked glycosylation, had not been translocated across the ER membrane. However, it was important to provide evidence that the function of the S/A domain had not been abrogated and

A.

MUTANT	CHARGE ^a		$\Delta(C-N)^b$		% N _{exo} ^c
	N	C			
—	+1	0	-1	MVAEDAPVRATCRVLFR  ESLITOKQIMSQAGSTG	0
1	-1	0	+1	— E-S  —	12
2	-2	0	+2	— E-E  —	30
3	+1	+2	+1	—  — R-R	5
4	+1	+3	+2	—  K-R	5
5	+1	+4	+3	—  K-R-R	6
6	-1	+3	+4	— E-S  K-R	20
7	-1	+4	+5	— E-S  K-R-R	50
8	-2	+3	+5	— E-E  K-R	50
9	-2	+4	+6	— E-E  K-R-R	75

B.

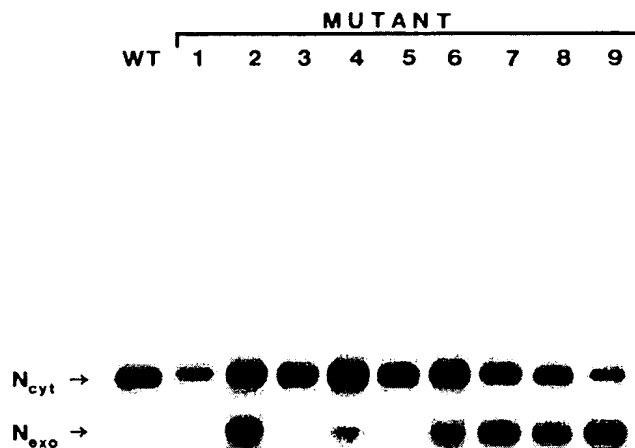


Figure 1. Construction and Expression of Charge-Altered HN Proteins

(A) Schematic diagram of the charge-altered HN proteins. The 17 amino acid residues flanking the N- (left) and C-terminal (right) sides of the S/A (cross-hatched box) of WT HN are shown. Solid horizontal lines denote sequence identity of mutants 1-9 with WT HN, and substitutions are shown below their position in the HN sequence. a: sum of charged residues within the 15 amino acids flanking the S/A domain; N, N-terminal; C, C-terminal. b: difference in the sum of charged residues on N- and C-terminal sides of S/A. c: percentage of the total HN protein accumulated in the unglycosylated N_{exo} form after a 1 hr labeling period.

(B) Expression of charge-altered HN proteins. CV-1 cells infected with vaccinia virus vTF7-3 were transfected with plasmids encoding WT HN or mutants 1-9 and radiolabeled for 1 hr with Tran^[35S]label. Proteins were immunoprecipitated from cell lysates with HN antisera and analyzed by SDS-PAGE. N_{cyt} and N_{exo} denote forms of HN as described in the text.



Figure 2. Biochemical Analysis of Microsomal Membranes from Cells Expressing Charge-Altered HN Proteins

Vaccinia virus vTF7-3-infected cells were transfected with plasmids encoding WT HN or with mutants 2, 6, or 7. Cells were radiolabeled with Tran[³⁵S]label from 3.5–4.5 hr posttransfection, and crude microsomal membranes were prepared.

(A) Alkali fractionation. Microsomal membranes were incubated for 30 min at 4°C with buffer (pH 11) and fractionated by centrifugation. Equal portions of the resulting pellet (P) or supernatant (S) were neutralized, immunoprecipitated with HN antisera, and the polypeptides were analyzed by SDS-PAGE.

(B) Protease digestion. Samples were treated with buffer (– lanes) or with 20 µg/ml trypsin (+ lanes). After 45 min at 37°C, microsomal membranes were isolated by centrifugation, and the proteins were immunoprecipitated with HN antisera before analysis by SDS-PAGE. N_{cyt} and N_{exo} are forms of HN as described in the text.

that these unglycosylated molecules were stably anchored in the membrane (N_{exo}C_{cyt} orientation) and were not soluble cytoplasmic proteins. Microsomal membranes were prepared from vTF7-3-infected cells that had been transfected with plasmids encoding WT HN or mutants 2, 6, or 7, and the microsomes were treated with pH 11 buffer. Under these conditions, integral membrane proteins remain associated with the lipid bilayer and after centrifugation are found in the pellet fraction, while soluble proteins are found in the supernatant fraction (Steck and Yu, 1973). As shown in Figure 2A, both the N_{cyt} and the N_{exo} protein species fractionated like WT HN, as the majority of the protein was detected in the pellet fraction (P) and only trace amounts were found in the supernatant (S). Thus, these data strongly suggest that the function of the S/A domain in targeting the proteins to the ER and anchoring the proteins in membranes had not been affected.

To provide direct biochemical evidence for the topology of the mutant proteins, microsomal membranes isolated from vaccinia vTF7-3-infected cells expressing WT HN or several representative mutants were treated with trypsin, and the protected protein fragments were analyzed by immunoprecipitation with HN antisera and SDS-PAGE. Microsomal membranes from cells expressing WT HN or

mutants 2 and 7 protected the N_{cyt} species from trypsin digestion, whereas the N_{exo} form was accessible to added protease (Figure 2B, + lanes). These results suggest that the N_{cyt} species has a type II orientation and that the vast majority of the N_{exo} polypeptide chain is located on the cytoplasmic side of the membrane.

To provide evidence that the N-terminal domain of the HN N_{exo} species was translocated across the ER membrane and not held in a loop formation, a site for the addition of N-linked glycosylation was added to the N-terminal domain of WT HN and two of the charge-altered mutants by site-specific mutagenesis (Figure 3). It was anticipated that glycosylation of the N-terminal domain of the N_{exo} species would result in a slower electrophoretic mobility than the unglycosylated N_{exo} protein, while the mobility of the N_{cyt} species would not be altered. Vaccinia virus vTF7-3-infected cells were transfected with plasmids encoding these N-terminal mutants and labeled for 1 hr with Tran[³⁵S]label. Proteins were immunoprecipitated from cell extracts, incubated with (+) or without (–) N-glycanase, and examined by SDS-PAGE. The mutant HN WT* contains the new N-terminal site for N-linked glycosylation, and expression of HN WT* results in the synthesis of a single major polypeptide (Figure 3, WT* lanes). Thus,

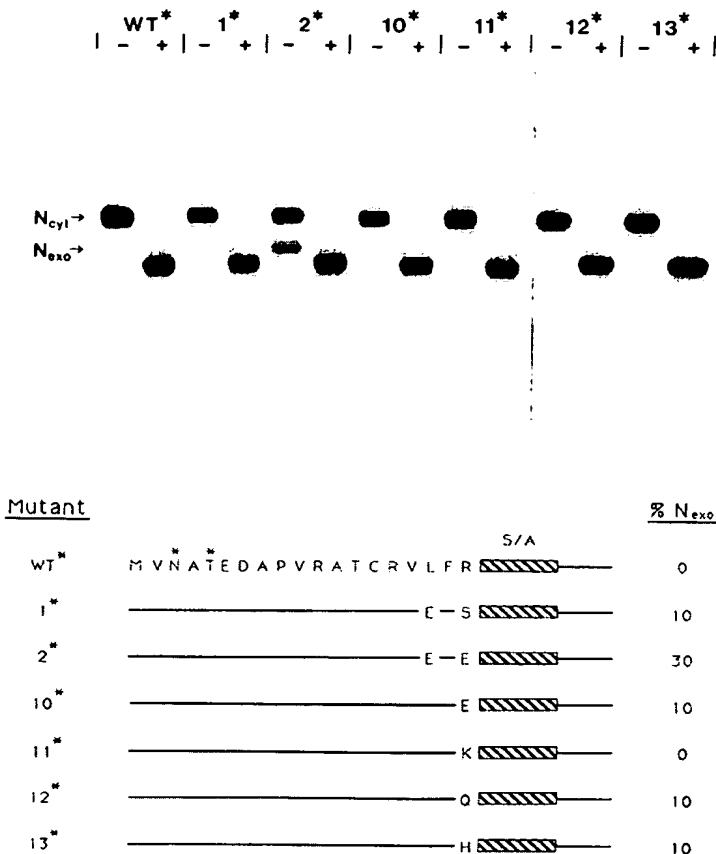


Figure 3. Glycosylation of the Mutant HN N-Terminal Domains

Vaccinia vTF7-3-infected CV-1 cells were transfected with plasmid DNAs encoding derivatives of the WT and mutant HN proteins altered to contain an N-terminal glycosylation site (*). Polypeptides were radiolabeled from 3.5–4.5 hr posttransfection with Tran[³⁵S]label and immunoprecipitated with HN antisera. Immune complexes were divided into two portions, incubated with (+) or without (–) N-glycanase, and the polypeptides were analyzed by SDS-PAGE. The fraction of the total HN protein in the N_{exo} orientation is shown (% N_{exo}). The N-terminal amino acids in the mutants are listed with solid horizontal lines, indicating sequence identity with HN WT*. Note that HN WT* contains two extra N-terminal residues (N and T) to create the site for N-linked glycosylation. S/A, HN signal/anchor domain.

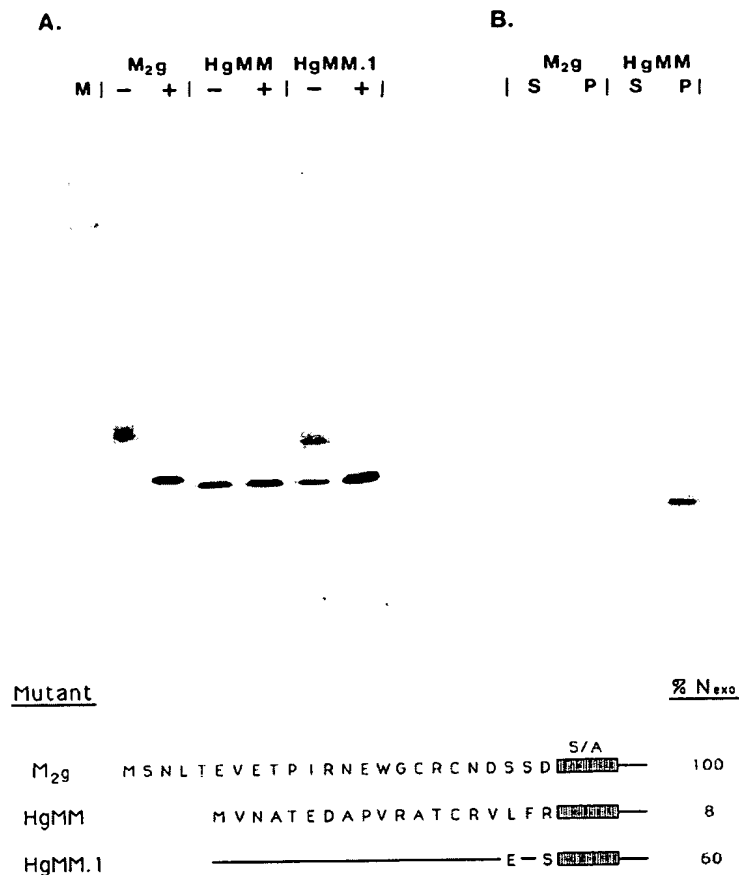
the addition of the two new amino acid residues to form the N-terminal glycosylation site did not influence HN orientation. Two polypeptide species were identified with HN mutants 1* and 2* (– lanes, N_{cyl} and N_{exo}), both of which had a slower electrophoretic mobility than the single polypeptide species found after treatment of the proteins with N-glycanase (+ lanes). The small mobility difference of ~5 kd between the N_{exo} species (– lanes) and the deglycosylated protein (+ lanes) suggested that the N_{exo} polypeptides had been modified by the addition of carbohydrate and the shift in mobility is consistent with the use of the new N-terminal glycosylation site. Furthermore, the relative abundance of the singly glycosylated 1* and 2* N_{exo} forms (10% and 30%) correlates well with the amount of their unglycosylated HN counterparts seen in Figure 1 (12% and 30%). Taken together, these biochemical data indicate that the mutant N_{exo} species represents an integral membrane protein with a large C-terminal cytoplasmic region and a small N-terminal domain in the ER lumen, and thus these molecules are the result of a bona-fide inversion of the HN type II topology.

Additional HN mutants (Figure 3, 10*–13*) were constructed to determine whether an arginine (R) residue directly flanking the HN N-terminal side of the S/A was re-

quired for the establishment of the N_{cyl}C_{exo} topology. The HN WT* cDNA was altered by mutagenesis such that a negatively charged glutamic acid (E), a positively charged lysine (K), an uncharged glutamine (Q), or a histidine (H) residue, the latter which can be weakly positively charged depending on the intracellular pH, replaced the R residue that normally flanks the S/A domain. These mutants were expressed from plasmids and analyzed as described above for mutants HN WT*, 1*, and 2*. As shown in Figure 3, approximately 10% of the mutants containing the E, Q, or H substitution were found in the N_{exo} form, and this value was very similar to that obtained with the 1* mutant. Expression of the K substitution mutant 11* (lane 11*) resulted in a polypeptide mobility pattern that was indistinguishable from that of HN WT*, indicating that a positively charged residue directly flanking the N-terminal side of the S/A is very important for establishing the HN type II topology.

The HN N-Terminal Domain Directs the Inversion of M₂ Into a Type II Topology

The above data suggest that the HN N-terminal domain plays a critical role in governing the type II orientation and that this region may contain a signal that is incompatible



with its translocation across the membrane. A prediction of this proposal would be that the type III N_{exo}C_{cyt}-oriented M₂ protein would lack this N-terminal retention signal, but that transfer of the HN N-terminal domain to the M₂ protein should direct an inversion of M₂ to the type II topology. To test this prediction, a hybrid cDNA molecule was constructed (Figure 4, HgMM) such that it encoded the HN WT* N-terminal 19 residues linked precisely to the M₂ S/A and cytoplasmic domains. The addition of carbohydrate residues to this chimera would indicate that the HN N-terminal domain, which contains the only site for N-linked glycosylation, has been translocated across the ER membrane. Vaccinia virus vTF7-3-infected cells were transfected with plasmids encoding the HgMM chimera or M₂g, a modified version of the M₂ protein that contains an N-terminal site for N-linked glycosylation to facilitate the analysis of M₂ membrane topology (Parks et al., 1989). The cells were labeled with [³⁵S]cysteine and [³⁵S]methionine for 2 hr, and the proteins were immunoprecipitated with M₂-specific antisera, incubated with (+) or without (-) N-glycanase, and examined by SDS-PAGE.

The M₂g protein was synthesized as a major species (Figure 4, M₂g, - lane), which has a slower elec-

trophoretic mobility than the N-glycanase-treated protein (+ lane); this is consistent with the known N_{exo}C_{cyt} topology of M₂ (Lamb et al., 1985). The M₂g protein was observed to migrate as a doublet; this may reflect differential processing of the carbohydrate residues. In contrast, only 8% of the HgMM protein was glycosylated and the vast majority of the HgMM protein was synthesized as an unglycosylated polypeptide (HgMM, - lane) exhibiting an electrophoretic mobility indistinguishable to that of the N-glycanase-treated sample (+ lane). Alkali extraction of microsomal membranes isolated from cells expressing M₂g or HgMM showed that both of these polypeptides were strongly associated with the membrane, as they were found in the pellet fraction and not in the supernatant (Figure 4B). Thus, these data indicate that the vast majority of the chimeric HgMM protein is orientated as a type II protein. Parenthetically, the observed type II topology of HgMM differs from that predicted by the charge difference rules (Hartmann et al., 1989), as the sum of the charges flanking the HgMM S/A on the N- and C-terminal sides are +1 and +1.5, respectively.

The results obtained with HN mutants 10* -13* indicate that a positive charge immediately flanking the HN S/A is

Figure 4. Transfer of the HN N-Terminal Domain to M₂ Results in a Type II Chimeric Polypeptide

(A) Expression of M₂g, HgMM, and HgMM.1. Vaccinia vTF7-3-infected CV-1 cells were transfected with plasmids encoding M₂g, HgMM, or HgMM.1 and radiolabeled for 2 hr with [³⁵S]methionine and [³⁵S]cysteine. Samples were immunoprecipitated from cell extracts with M₂ antisera, incubated with (+) or without (-) N-glycanase and analyzed by SDS-PAGE. Lane M, influenza A virus-infected cell polypeptides as a marker; the fastest-migrating species is authentic M₂ polypeptide.

(B) Alkali extraction of membranes from cells expressing M₂g or HgMM. Vaccinia virus vTF7-3-infected cells were transfected with plasmids encoding M₂g or HgMM and were radiolabeled for 2 hr with [³⁵S]methionine and [³⁵S]cysteine. Crude microsomal membranes were prepared, incubated with pH 11.0 buffer, and fractionated by centrifugation. Equal portions of the resulting supernatant (S) or pellet (P) fractions were immunoprecipitated with M₂ antisera, and the samples were examined by SDS-PAGE. The N-terminal amino acid sequences of the mutants are listed with the HgMM.1 N-terminal horizontal line denoting sequence identity with HgMM (Hg is identical to HN WT*). S/A, M₂ signal/anchor domain.

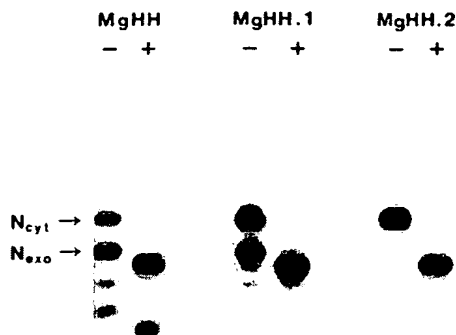


Figure 5. Positive N-Terminal Charges Convert the MgHH Chimeric Protein to a Type II Orientation

CV-1 cells infected with vaccinia vTF7-3 were transfected with plasmid DNA encoding MgHH, MgHH.1, or MgHH.2 and radiolabeled for 1 hr with Tran³⁵S]label. HN proteins were immunoprecipitated from cell extracts with HN antisera, incubated with (+) or without (-) N-glycanase, and the polypeptides were examined by SDS-PAGE. The N-terminal amino acids in the mutants are listed with horizontal lines denoting sequence identity with M₂g. S/A, HN signal/anchor domain.

Mutant		% N _{exo}
MgHH	MSNLTEVETPIRNEWGCRCDSSD S/A	60
MgHH.1	-----R-----	40
MgHH.2	-----G-----R-----	3

important for establishing a type II orientation. A charge-altered form of the HgMM chimera (HgMM.1, Figure 4) was constructed to test whether a positive charge flanking the N-terminal side of the S/A was also a factor in establishing the HgMM N_{cyt}C_{exo} topology. The HgMM.1 mutant, which encoded the same L to E and R to S mutations previously analyzed in HN mutant 1*, was expressed and analyzed as described above for M₂g and HgMM. As shown in Figure 4, this charge-altered chimera (HgMM.1, - lane) was synthesized as a mixture of a slow-migrating glycosylated form and a faster-migrating species with a mobility matching that of the N-glycanase-treated sample (+ lane). In contrast to the predominant type II orientation of the unaltered HgMM hybrid, approximately 60% of the modified HgMM.1 protein was found to be glycosylated and thus must be in an N_{exo}C_{cyt} topology. Thus, these data suggest that the HN N-terminal region can direct an inversion of the M₂ polypeptide from the N_{exo}C_{cyt} topology to that of a type II protein, and this efficient inversion is disrupted by the removal of N-terminal positively charged residues.

Ability of the M₂ N-Terminal Domain to Direct the Type II Topology

We were interested in determining if the reciprocal experiment to that described in the section above could be performed, i.e., to convert a type III N_{exo} domain into a type II N_{cyt} domain by the addition of positively charged residues. We have previously described the properties of

a chimeric M₂/HN protein containing the M₂g N-terminal 24 residues linked precisely to the HN S/A and C-terminal domains (Parks et al., 1989). This MgHH hybrid is synthesized as a single polypeptide chain that adopts two opposing orientations in membranes, with approximately 60% of the protein in the faster-migrating N_{exo} form (Figure 5, MgHH panel). Minor faster-migrating species are degradation products of the HN ectodomain (Parks et al., 1989; Ng et al., 1989). The effect of the addition of positively charged N-terminal residues on the orientation of this hybrid was examined by constructing two charge-altered MgHH mutants.

In MgHH.1, a single R residue was substituted for the M₂g N-terminal serine at amino acid residue 23, and MgHH.2 encoded a substitution of the two negatively charged M₂ aspartic acid (D) residues at positions 21 and 24 with glycine (G) and arginine (R), respectively (Figure 5). The rationale for the addition of a G residue at position 21 in MgHH.2 was based on the finding that this was a naturally occurring change in the N-terminal ectodomain between the M₂ proteins of the Udorn/72 and PR/8/34 strains of Influenza A virus (Lamb and Lai, 1981; Lamb et al., 1985). Thus, it is known that a D to G substitution at this position does not alter the M₂ protein orientation but would contribute generally to the N-terminal charge distribution. CV-1 cells infected with vaccinia vTF7-3 were transfected with plasmids encoding the altered MgHH hybrids and labeled for 1 hr with Tran³⁵S]label. Proteins were immunoprecipitated with HN antisera, incubated with (+) or

without (–) N-glycanase, and analyzed by SDS–PAGE. As shown in Figure 5, the MgHH.1 mutant was synthesized as two major species (Figure 5, panel MgHH.1, – lane) that were converted to a single faster-migrating form after N-glycanase treatment (+ lane), and 40% of this modified chimera was in the N_{exo} orientation. In contrast, the MgHH.2 mutant that contained a positively charged arginine residue flanking the S/A was predominantly in the type II N_{Cyt} orientation, with only 3% of the protein in the N_{exo} topology (Figure 5, panel MgHH.2, – lane). Thus, these data indicate that the addition of positively charged residues to the M₂ N-terminal ectodomain next to the S/A domain alters this region such that it can adopt a type II topology.

Discussion

We wished to test the role of charged residues flanking the S/A domain in determining orientation since the biochemical mechanism involved in generating the topology of eukaryotic membrane proteins with an internal uncleaved S/A has not been established. For the purposes of discussion the boundaries of the S/A domain are defined as the first charged residue in both directions from the middle of the first hydrophobic domain. The signals directing the orientation of proteins in the ER membrane can be thought of in simple terms as either acting to promote the translocation of the N-terminus of a type III protein across the membrane, acting to retain the N-terminus of type II proteins in the cytoplasm, or both signals could exist with one being dominant. Our data emphasize the importance of N-terminal positive charges in generating the type II orientation. Removal of positively charged residues from the N_{Cyt} domain resulted in some of the HN molecules assuming an inverted orientation in membranes. However, as the inversion was not absolute it suggests that the absence of a positively charged residue is not the sole factor involved in generating the type III orientation. In part, the mixed orientation of the chimeric proteins (i.e., MgHH) before the charges were altered may reflect difficulties involved with using chimeric proteins rather than naturally existing proteins. Interestingly, the addition of charges to the C-terminal side of the HN S/A domain in the absence of the N-terminal positive charge residue resulted in more efficient inversion as discussed further below. Previously it has been found that the addition of N-terminal positively charged residues inverts the type III cytochrome P₄₅₀ protein but because of exposure of a cryptic site for cleavage by signal peptidase it becomes a secreted protein (Szczesna-Skorupa et al., 1988; Sato et al., 1990). In addition, it has been found that by switching domains in chimeric proteins, which leads to alterations in the positions of charged residues, membrane topology can be altered both *in vitro* and *in vivo* (Haeuptle et al., 1989; Parks et al., 1989).

We favor the idea that the N-terminal positively charged residue flanking the S/A domain is an important part of a dominantly acting retention signal that retains the N-terminus of the nascent polypeptide chain in the cytoplasm to create the type II orientation, and that this retention sig-

nal is not present in the N-terminal domain of type III proteins. This conclusion is based on several lines of evidence in addition to the data obtained with the N-terminal charge-altered mutants. First, linking of the HN N-terminal domain to the M₂ S/A and C-terminal regions produces a chimeric protein (HgMM) that largely adopts the HN topology, indicating that the dominant determinant of type II topology had been transferred to M₂, and that this HN signal could efficiently override any possible topological signals in the M₂ S/A and cytoplasmic domains. Second, the M₂ N-terminal ectodomain although only 60% efficient at directing the chimera MgHH into the type III orientation can be altered to efficiently direct the MgHH chimera into the type II orientation when positive charges are introduced into the N-terminal S/A flanking positions. This suggests that the nature of the sequence that comprises a cytoplasmic domain is less critical for generating type II topology than the presence of the appropriately positioned positively charged residue, and that it is possible to create the signal that specifies type II topology.

These data support the "positive inside" rule proposed previously (von Heijne and Gavel, 1988) and for which evidence has recently been provided in the case of a bacterial membrane protein (Nilsson and von Heijne, 1990) in that positive charges are an important factor directing HN membrane topology. However, the orientation of the HN protein is more sensitive to the removal of N-terminal positive charges than to the addition of C-terminal positive charges, and this indicates that the topology of eukaryotic type II proteins is not determined simply by the retention of the most positively charged domain. Once the N-terminal positive charge has been removed, the subsequent addition of positive charges to the C-terminal side of the S/A may operate to keep this domain in the cytoplasm (Figure 1, mutants 6–9). Thus, eukaryotes and prokaryotes may share a common mechanism for generating membrane protein topology by which charged residues provide a barrier to translocation, but their mechanisms may differ from each other in the relative importance of N-terminal positive charges.

It was originally suggested on theoretical grounds, and then supported experimentally, that the signal sequence of type I and II proteins is inserted into the ER membrane as a hairpin loop with both the N- and C-terminal regions located in the cytoplasm (von Heijne and Blomberg, 1979; Inouye and Halegoua, 1980; Engelman and Steitz, 1981; Shaw et al., 1988). As the insertion of type III membrane proteins into membranes is dependent on recognition of the S/A by the signal recognition particle (Hull et al., 1988), the nascent type III chain probably also adopts a loop structure. However, after membrane insertion as a hairpin loop, the critical step in generating the type III topology involves the translocation of the N-terminal domain across the lipid bilayer. The N- to C-terminal polarity of protein synthesis implies that the N-terminal region flanking the S/A of a nascent polypeptide would be exposed to the translocation machinery prior to complete exposure of the C-terminal flanking region, and it has been suggested that the transfer of the type III N-terminal domain across the membrane may occur very fast relative to the rate of

translation (von Heijne, 1986b). In contrast, the presence of a positive-charge signal in the N-terminal region of the nascent polypeptide chain of type I proteins and the mature polypeptide chain of type II proteins imparts cytoplasmic retention of this domain and the C-terminal region is translocated. Although the topology of the immature type I and mature type II proteins may ultimately be dependent on the presence or absence of an available site for cleavage by signal peptidase (Lipp and Dobberstein, 1986a; Shaw et al., 1988), what distinguishes them from type III proteins is that during synthesis there is retention of the N-terminus in the cytoplasm.

It is not known whether retention of the N-terminal domain of nascent type I and mature type II proteins is due to binding by cytoplasmic factors or if a local electrical potential across the membrane makes translocation of this region thermodynamically unfavorable (Weinstein et al., 1982). The translocation of a polypeptide chain into the ER could occur, at least in theory, by direct transfer through the hydrophobic environment of the lipid bilayer (von Heijne and Blomberg, 1979; Engelman and Steitz, 1981) or through a protein pore in the membrane (Blobel and Dobberstein, 1975; Gilmore and Blobel, 1985), but recent evidence suggests that during translocation the nascent chain is associated with distinct membrane-bound proteins (Connolly et al., 1989; Nicchitta and Blobel, 1989). In the case of prokaryotes, it has been suggested that the *Escherichia coli* SecA protein directs protein translocation by recognizing N-terminal positive charges within a signal sequence (Akita et al., 1990), and it seems possible that an analogous protein may operate similarly in eukaryotes. The ability to reconstitute membrane translocation in vitro from disrupted microsomes (Nicchitta and Blobel, 1990) may provide the means to separate and assess the role of individual components of the translocation machinery in directing membrane topology.

Experimental Procedures

Plasmid Construction and Site-Specific Mutagenesis
cDNA clones that express wild-type SV5 HN (pSV103HNm, Hiebert et al., 1985a; Paterson et al., 1985) and M₂g, a derivative of influenza A virus M₂ containing an N-terminal site for N-linked glycosylation (Parks et al., 1989), were used as a source of starting materials for the construction of the altered genes. Bacteriophage M13M₂g (containing the entire M₂g cDNA in the BamHI site of the replicative form of M13mp19) and M13HN (containing 5' nucleotides 1–306 and encoding N-terminal residues 1–81 from the HN gene) were used as template DNA for oligonucleotide-directed mutagenesis as described previously (Parks et al., 1989). Oligonucleotides were synthesized by the Northwestern University Biotechnology Facility on a DNA synthesizer (Model 380B, Applied Biosystems Inc., Foster City, CA).

Mutant HN DNA segments were excised from the replicative form of M13 by EcoRI and PstI digestion, subcloned into a pGem vector, and their nucleotide sequence confirmed by dideoxynucleotide chain-terminating sequencing (Sanger et al., 1977). The altered 5' end DNA fragments were then reconstructed into a full-length gene by linkage to the HN PstI–XhoI fragment (encoding residues 82–565) in pGem3 such that mRNA sense transcripts could be generated using the bacteriophage T₇ RNA polymerase promoter and T₇ RNA polymerase. pGem-HNWT*, which encodes an N-terminal site for N-linked glycosylation (Asn-Ala-Thr), was constructed by the insertion of codons for Asn and Thr between HN bases 72–73 and 75–76, respectively.

The HgMM gene was constructed by introducing a new StuI site (AGGCCT), which encodes the junction of the HN N-terminal and M₂

S/A domains (Arg-Pro), into the HN WT* (bases 115–120) and M₂ (bases 95–100) cDNA fragments by oligonucleotide-directed mutagenesis. Blunt-end ligation of the EcoRI–StuI HN WT* fragment to the StuI–PstI M₂ fragment in the EcoRI and PstI sites of pGem3 yielded a DNA segment that encoded the HN WT* N-terminal residues 1–19 linked precisely to the M₂ S/A and C-terminal domains (residues 25–52, Lamb et al., 1985). Similarly, HgMM.1 was constructed by blunt-end ligation of the HN mutant 1* EcoRI–ScaI and M₂ StuI–PstI fragments into pGem3. The construction and characterization of MgHH has been described previously (Parks et al., 1989). MgHH.1 and MgHH.2 were constructed by site-specific mutagenesis as described (Parks et al., 1989). Nucleotide sequences were confirmed by dideoxynucleotide chain-terminating sequencing (Sanger et al., 1977).

Cells

Monolayer cultures of CV-1 cells were grown in Dulbecco's modified Eagle's medium containing 10% fetal calf serum as described (Lamb and Lai, 1982).

Isotopic Labeling of Polypeptides, Immunoprecipitation, Peptide:N-Glycosidase F Digestions, and Polyacrylamide Gel Electrophoresis

cDNA clones were expressed by a modification of the vaccinia virus/T₇ RNA polymerase system of Fuerst et al. (1986). In brief, confluent monolayers of CV-1 cells (6 cm diameter plates) were infected (multiplicity of infection = 20) with recombinant vaccinia virus vTF7-3, which encodes the bacteriophage T₇ RNA polymerase. The inoculum was removed and calcium phosphate-precipitated plasmid DNA (~30 µg) was then added. Cells transfected with plasmids encoding HN mutants were radiolabeled from 3.5–4.5 hr posttransfection with 20–50 µCi/ml Tran[³⁵S]label (ICN Radiochemicals Inc., Irvine, CA) in Dulbecco's modified Eagle's medium lacking cysteine and methionine. Radiolabeled cells were washed in phosphate-buffered saline and lysed in 1% SDS. Immunoprecipitation from cell extracts using polyclonal rabbit sera to denatured HN (HN antisera) was as described (Ng et al., 1990; Erickson and Blobel, 1979). Cells transfected with plasmids encoding M₂g or the HN/M₂ hybrids were radiolabeled from 3.5–5.5 hr posttransfection with a mixture of [³⁵S]cysteine and [³⁵S]methionine (125 µCi/ml each), and the proteins were immunoprecipitated from cells solubilized in cold RIPA buffer (Lamb et al., 1978) using polyclonal sera raised to denatured M₂ (DM2 antisera, Lamb et al., 1985). Treatment of samples with peptide:N-glycosidase F was as described previously (Williams and Lamb, 1986). Samples were analyzed by SDS–PAGE on 10% polyacrylamide gels (HN proteins) or 17.5% gels containing 4 M urea (M₂g and HN/M₂ hybrid proteins), followed by fluorography (Lamb and Chopin, 1976). Densitometric scanning of autoradiograms was carried out using an LKB Ultrascan XL laser densitometer (Pharmacia-LKB, Bromma, Sweden). The %N_{total} values reported represent the average of at least two experiments.

Trypsin Digestion and Alkali Extraction of Microsomal Membranes

Vaccinia virus vTF7-3-infected cells were transfected with plasmid DNAs and radiolabeled from 3.5–4.5 hr post-DNA transfection with 20 µCi/ml Tran[³⁵S]label (HN proteins) or from 3 to 5 hr posttransfection with 250 µCi/ml [³⁵S]cysteine and [³⁵S]methionine (M₂g and HN/M₂ proteins) before the preparation of crude microsomal membranes by Dounce homogenization (Adams and Rose, 1985). Samples were analyzed by trypsin digestion or alkali fractionation as described previously (Parks et al., 1989).

Acknowledgments

We thank Margaret Shaughnessy for excellent technical assistance and David Simpson for advice on the vaccinia virus expression system. We thank Bernard Moss and Thomas Fuerst for supplying vaccinia virus vTF7-3. G. D. P. was supported by an American Cancer Society postdoctoral fellowship (PF-3177). This research was supported by Public Health Service research grants AI-20201 and AI-23173 from the National Institute of Allergy and Infectious Diseases.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby

marked "advertisement" in accordance with 18 USC Section 1734 solely to indicate this fact.

Received October 17, 1990; revised November 19, 1990.

References

- Adams, G. A., and Rose, J. K. (1985). Incorporation of a charged amino acid into the membrane spanning domain blocks cell surface transport but not membrane anchoring of a viral glycoprotein. *Mol. Cell. Biol.* 5, 1442-1448.
- Akita, M., Sasaki, S., Matsuyama, S.-i., and Mizushima, S. (1990). SecA interacts with secretory proteins by recognizing the positive charge at the amino terminus of the signal peptide in *Escherichia coli*. *J. Biol. Chem.* 265, 8164-8169.
- Blobel, G. (1990). Intracellular protein topogenesis. *Proc. Natl. Acad. Sci. USA* 77, 1496-1500.
- Blobel, G., and Dobberstein, B. (1975). Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J. Cell Biol.* 67, 835-851.
- Connolly, T., Collins, P., and Gilmore, R. (1989). Access of proteinase K to partially translocated nascent polypeptides in intact and detergent solubilized membranes. *J. Cell Biol.* 108, 299-308.
- Engelman, D. M., and Steitz, T. A. (1981). The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. *Cell* 23, 411-422.
- Erickson, A. H., and Blobel, G. (1979). Early events in the biosynthesis of lysosomal enzyme cathepsin. *J. Biol. Chem.* 254, 11771-11774.
- Evans, E. A., Gilmore, R., and Blobel, G. (1986). Purification of microsomal signal peptidase as a complex. *Proc. Natl. Acad. Sci. USA* 83, 581-585.
- Fuerst, T. R., Niles, E. G., Studier, F. W., and Moss, B. (1986). Eukaryotic transient-expression system based on recombinant vaccinia virus that synthesizes bacteriophage T₇ RNA polymerase. *Proc. Natl. Acad. Sci. USA* 83, 8122-8126.
- Gilmore, R., and Blobel, G. (1985). Translocation of secretory proteins across the microsomal membrane occurs through an environment accessible to aqueous perturbants. *Cell* 42, 497-505.
- Haeuptle, M.-T., Flint, N., Gough, N. M., and Dobberstein, B. (1989). A tripartite structure of the signals that determine protein insertion into the endoplasmic reticulum membrane. *J. Cell Biol.* 108, 1227-1236.
- Hartmann, E., Rapoport, T. A., and Lodish, H. F. (1989). Predicting the orientation of eukaryotic membrane-spanning proteins. *Proc. Natl. Acad. Sci. USA* 86, 5786-5790.
- Hiebert, S. W., Paterson, R. G., and Lamb, R. A. (1985a). Hemagglutinin-neuraminidase protein of the paramyxovirus simian virus 5: nucleotide sequence of the mRNA predicts an N-terminal membrane anchor. *J. Virol.* 54, 1-6.
- Hiebert, S. W., Paterson, R. G., and Lamb, R. A. (1985b). Identification and predicted sequence of a previously unrecognized small hydrophobic protein, SH, of the paramyxovirus simian virus 5. *J. Virol.* 55, 744-751.
- High, S., and Tanner, M. J. A. (1987). Human erythrocyte membrane sialoglycoprotein β . *Biochem. J.* 243, 277-280.
- Hull, J. D., Gilmore, R., and Lamb, R. A. (1988). Integration of a small integral membrane protein, M₂, of influenza virus into the endoplasmic reticulum: analysis of the internal signal-anchor domain of a protein with an ectoplasmic NH₂ terminus. *J. Cell Biol.* 106, 1489-1498.
- Inouye, M., and Halegoua, S. (1980). Secretion and membrane localization of proteins in *Escherichia coli*. *CRC Crit. Rev. Biochem.* 7, 339-371.
- Lamb, R. A., and Chopin, P. W. (1976). Synthesis of influenza virus proteins in infected cells: translation of viral polypeptides, including three P polypeptides, from RNA produced by primary transcription. *Virology* 74, 504-519.
- Lamb, R. A., and Lai, C.-J. (1981). Conservation of the influenza virus membrane protein (M₁) amino acid sequence and an open reading frame of RNA segment 7 encoding a second protein (M₂) in H1N1 and H3N2 strains. *Virology* 112, 746-751.
- Lamb, R. A., and Lai, C.-J. (1982). Spliced and unspliced messenger RNAs synthesized from cloned influenza virus DNA in an SV40 vector: expression of the influenza virus membrane protein (M₁). *Virology* 123, 237-256.
- Lamb, R. A., Etchick, P. R., and Chopin, P. W. (1978). Evidence for a ninth influenza viral polypeptide. *Virology* 91, 60-78.
- Lamb, R. A., Zebadee, S. L., and Richardson, C. D. (1985). Influenza virus M₂ protein is an integral membrane protein expressed on the infected-cell surface. *Cell* 40, 627-633.
- Lipp, J., and Dobberstein, B. (1986a). The membrane-spanning segment of invariant chain (h) contains a potentially cleavable signal sequence. *Cell* 46, 1103-1112.
- Lipp, J., and Dobberstein, B. (1986b). Signal recognition particle-dependent membrane insertion of mouse invariant chain: a membrane-spanning protein with a cytoplasmically exposed amino terminus. *J. Cell Biol.* 102, 2169-2175.
- Ng, D. T. W., Randall, R. E., and Lamb, R. A. (1989). Intracellular maturation and transport of the SV5 type II glycoprotein hemagglutinin-neuraminidase: specific and transient association with GRP78-BiP in the endoplasmic reticulum and extensive internalization from the cell surface. *J. Cell Biol.* 109, 3273-3289.
- Ng, D. T. W., Hiebert, S. W., and Lamb, R. A. (1990). Different roles of individual N-linked oligosaccharide chains on the folding, assembly and transport of the simian virus 5 hemagglutinin-neuraminidase integral membrane glycoprotein. *Mol. Cell. Biol.* 10, 1989-2001.
- Nicchitta, C. V., and Blobel, G. (1989). Nascent secretory chain binding and translocation are distinct processes: differentiation by chemical alkylation. *J. Cell Biol.* 108, 789-796.
- Nicchitta, C. V., and Blobel, G. (1990). Assembly of translocation-competent proteoliposomes from detergent-solubilized rough microsomes. *Cell* 60, 259-269.
- Nilsson, I., and von Heijne, G. (1990). Fine-tuning the topology of a polytopic membrane protein: role of positively and negatively charged amino acids. *Cell* 62, 1135-1141.
- Parks, G. D., Hull, J. D., and Lamb, R. A. (1989). Transposition of domains between the M₂ and HN viral membrane proteins results in polypeptides which can adopt more than one membrane orientation. *J. Cell Biol.* 109, 2023-2032.
- Paterson, R. G., Hiebert, S. W., and Lamb, R. A. (1985). Expression at the cell surface of biologically active fusion and hemagglutinin-neuraminidase proteins of the paramyxovirus SV5 from cloned cDNA. *Proc. Natl. Acad. Sci. USA* 82, 7520-7524.
- Sanger, F., Nicklin, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
- Sato, T., Sakaguchi, M., Mihara, K., and Omura, T. (1990). The amino-terminal structures that determine topological orientation of cytochrome P-450 in microsomal membranes. *EMBO J.* 9, 2391-2397.
- Schatzman, R. C., Evan, G. I., Privalsky, M. L., and Bishop, J. M. (1986). Orientation of the v-erb-B gene product in the plasma membrane. *Mol. Cell. Biol.* 6, 1329-1333.
- Schneider, C., Owen, M. J., Banville, D., and Williams, G. W. (1984). Primary structure of human transferrin receptor deduced from the mRNA sequence. *Nature* 311, 675-678.
- Shaw, A. S., Rottier, P. J. M., and Rose, J. K. (1988). Evidence for the loop model of signal-sequence insertion into the endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA* 85, 7592-7596.
- Spies, M., and Lodish, H. F. (1985). Sequence of a second human asialoglycoprotein receptor: conservation of two receptor genes during evolution. *Proc. Natl. Acad. Sci. USA* 82, 6465-6469.
- Steck, T. L., and Yu, J. (1973). Selective solubilization of proteins from red blood cell membranes by protein perturbants. *J. Supramol. Struct.* 1, 220-248.
- Strubin, M., Mach, B., and Long, E. O. (1984). The complete sequence of the mRNA for the HLA-DR-associated invariant chain reveals a polypeptide with an unusual transmembrane polarity. *EMBO J.* 3, 869-872.

- Szczesna-Skorupa, E., Browne, N., Mead, D., and Kemper, B. (1988). Positive charges at the N-terminus convert the membrane-anchor signal peptide of cytochrome P-450 to a secretory signal peptide. *Proc. Natl. Acad. Sci. USA* 85, 738-742.
- von Heijne, G. (1986a). The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* 5, 3021-3027.
- von Heijne, G. (1986b). Towards a comparative anatomy of N-terminal topogenic protein sequences. *J. Mol. Biol.* 189, 239-242.
- von Heijne, G. (1988). Transcending the impenetrable: how proteins come to terms with membranes. *Biochim. Biophys. Acta* 947, 307-333.
- von Heijne, G., and Blomberg, C. (1979). Trans-membrane translocation of proteins: the direct transfer model. *Eur. J. Biochem.* 97, 175-181.
- von Heijne, G., and Gavel, Y. (1988). Topogenic signals in integral membrane proteins. *Eur. J. Biochem.* 174, 671-678.
- Walter, P., and Lingappa, V. R. (1986). Mechanism of protein translocation across the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* 2, 499-516.
- Weinstein, J. N., Blumenthal, R., van Renswoude, J., Kempf, C., and Klausner, R. D. (1982). Charge clusters and the orientation of membrane proteins. *J. Membr. Biol.* 66, 203-212.
- Williams, M. A., and Lamb, R. A. (1986). Determination of the orientation of an integral membrane protein and sites of glycosylation by oligonucleotide-directed mutagenesis: influenza B virus NB glycoprotein lacks a cleavable signal sequence and has an extracellular N-terminal region. *Mol. Cell. Biol.* 6, 4317-4328.
- Zerial, M., Huylebroeck, D., and Garoff, H. (1987). Foreign transmembrane peptides replacing the internal signal sequence of transferrin receptor allow its translocation and membrane binding. *Cell* 48, 147-155.

Exhibit 28

Improved tools for biological sequence comparison

(amino acid/nucleic acid/data base searches/local similarity)

WILLIAM R. PEARSON* AND DAVID J. LIPMAN†

*Department of Biochemistry, University of Virginia, Charlottesville, VA 22908; and †Mathematical Research Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892

Communicated by Gerald M. Rubin, December 2, 1987 (received for review September 17, 1987)

ABSTRACT We have developed three computer programs for comparisons of protein and DNA sequences. They can be used to search sequence data bases, evaluate similarity scores, and identify periodic structures based on local sequence similarity. The FASTA program is a more sensitive derivative of the FASTP program, which can be used to search protein or DNA sequence data bases and can compare a protein sequence to a DNA sequence data base by translating the DNA data base as it is searched. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. The RDF2 program can be used to evaluate the significance of similarity scores using a shuffling method that preserves local sequence composition. The LFASTA program can display all the regions of local similarity between two sequences with scores greater than a threshold, using the same scoring parameters and a similar alignment algorithm; these local similarities can be displayed as a "graphic matrix" plot or as individual alignments. In addition, these programs have been generalized to allow comparison of DNA or protein sequences based on a variety of alternative scoring matrices.

We have been developing tools for the analysis of protein and DNA sequence similarity that achieve a balance of sensitivity and selectivity on the one hand and speed and memory requirements on the other. Three years ago, we described the FASTP program for searching amino acid sequence data bases (1), which uses a rapid technique for finding identities shared between two sequences and exploits the biological constraints on molecular evolution. FASTP has decreased the time required to search the National Biomedical Research Foundation (NBRF) protein sequence data base by more than two orders of magnitude and has been used by many investigators to find biologically significant similarities to newly sequenced proteins. There is a trade-off between sensitivity and selectivity in biological sequence comparison: methods that can detect more distantly related sequences (increased sensitivity) frequently increase the similarity scores of unrelated sequences (decreased selectivity). In this paper we describe a new version of FASTP, FASTA, which uses an improved algorithm that increases sensitivity with a small loss of selectivity and a negligible decrease in speed. We have also developed a related program, LFASTA, for local similarity analyses of DNA or amino acid sequences. These programs run on commonly available microcomputers as well as on larger machines.

METHODS

The search algorithm we have developed proceeds through four steps in determining a score for pair-wise similarity.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

FASTP and FASTA achieve much of their speed and selectivity in the first step, by using a lookup table to locate all identities or groups of identities between two DNA or amino acid sequences during the first step of the comparison (2). The *ktup* parameter determines how many consecutive identities are required in a match. For example, if *ktup* = 4 for a DNA sequence comparison, only those identities that occur in a run of four consecutive matches are examined. In the first step, the 10 best diagonal regions are found using a simple formula based on the number of *ktup* matches and the distance between the matches without considering shorter runs of identities, conservative replacements, insertions, or deletions (1, 3).

In the second step of the comparison, we rescore these 10 regions using a scoring matrix that allows conservative replacements and runs of identities shorter than *ktup* to contribute to the similarity score. For protein sequences, this score is usually calculated using the PAM250 matrix (4), although scoring matrices based on the minimum number of base changes required for a replacement or on an alternative measure of similarity can also be used with FASTA. For each of these best diagonal regions, a subregion with maximal score is identified. We will refer to this region as the "initial region"; the best initial regions from Fig. 1A are shown in Fig. 1B.

The FASTP program uses the single best scoring initial region to characterize pair-wise similarity; the initial scores are used to rank the library sequences. FASTA goes one step further during a library search; it checks to see whether several initial regions may be joined together. Given the locations of the initial regions, their respective scores, and a "joining" penalty (analogous to a gap penalty), FASTA calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. FASTA uses the resulting score to rank the library sequences. We limit the degradation of selectivity by including in the optimization step only those initial regions whose scores are above a threshold. This process can be seen by comparing Fig. 1B with Fig. 1C. Fig. 1B shows the 10 highest scoring initial regions after rescoring with the PAM250 matrix; the best initial region reported by FASTP is marked with an asterisk. Fig. 1C shows an optimal subset of initial regions that can be joined to form a single alignment.

In the fourth step of the comparison, the highest scoring library sequences are aligned using a modification of the optimization method described by Needleman and Wunsch (5) and Smith and Waterman (6). This final comparison considers all possible alignments of the query and library sequence that fall within a band centered around the highest scoring initial region (Fig. 1D). With the FASTP program, optimization frequently improved the similarity scores of related sequences by factors of 2 or 3. Because FASTA calculates an initial similarity score based on an optimization of initial regions during the library search, the initial score is

Abbreviation: NBRF, National Biomedical Research Foundation.

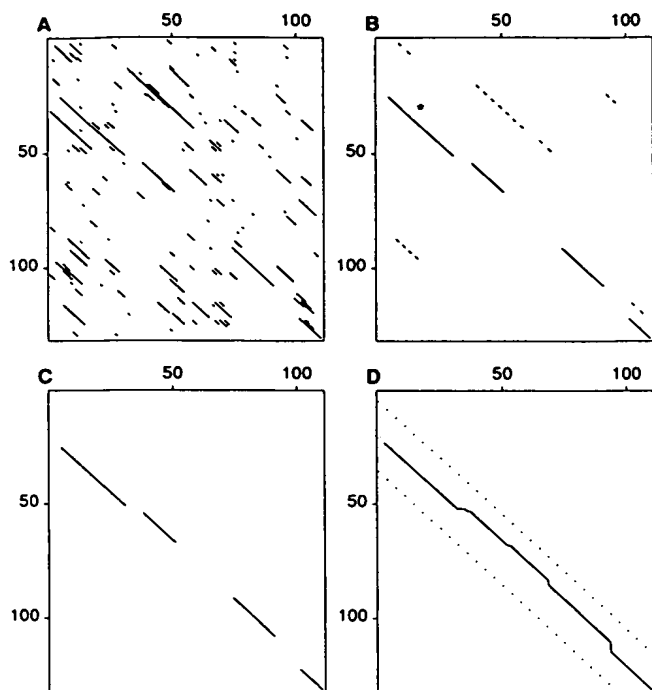


FIG. 1. Identification of sequence similarities by FASTA. The four steps used by the FASTA program to calculate the initial and optimal similarity scores between two sequences are shown. (A) Identify regions of identity. (B) Scan the regions using a scoring matrix and save the best initial regions. Initial regions with scores less than the joining threshold (27) are dashed. The asterisk denotes the highest scoring region reported by FASTP. (C) Optimally join initial regions with scores greater than a threshold. The solid lines denote regions that are joined to make up the optimized initial score. (D) Recalculate an optimized alignment centered around the highest scoring initial region. The dotted lines denote the bounds of the optimized alignment. The result of this alignment is reported as the optimized score.

much closer to the optimized score for many sequences. In fact, unlike FASTP, the FASTA method may yield initial scores that are higher than the corresponding optimized scores.

Local Similarity Analyses. Molecular biologists are often interested in the detection of similar subsequences within longer sequences. In contrast to FASTP and FASTA, which report only the one highest scoring alignment between two sequences, local sequence comparison tools can identify multiple alignments between smaller portions of two sequences. Local similarity searches can clearly show the results of gene duplications (see Fig. 2) or repeated structural features (see Fig. 3) and are frequently displayed using a "graphic matrix" plot (7), which allows one to detect regions of local similarity by eye. Optimal algorithms for sensitive local sequence comparison (6, 8, 9) can have tremendous computational requirements in time and memory, which make them impractical on microcomputers and, when comparing longer sequences, on larger machines as well.

The program for detecting local similarities, LFASTA, uses the same first two steps for finding initial regions that FASTA uses. However, instead of saving 10 initial regions, LFASTA saves all diagonal regions with similarity scores greater than a threshold. LFASTA and FASTA also differ in the construction of optimized alignments. Instead of focusing on a single region, LFASTA computes a local alignment for each initial region. Thus LFASTA considers all of the initial regions shown in Fig. 1B, instead of just the diagonal shown in Fig. 1D. Furthermore, LFASTA considers not

only the band around each initial region but also potential sequence alignments for some distance before and after the initial region. Starting at the end of the initial region, an optimization (6) proceeds in the reverse direction until all possible alignment scores have gone to zero. The location of the maximal local similarity score in the reverse direction is then used to start a second optimization that proceeds in the forward direction. An optimal path starting from the forward maximum is then displayed (5). The local homologies can be displayed as sequence alignments (see Fig. 2B) or on a two-dimensional graphic matrix style plot (see Figs. 2A and 3).

Statistical Significance. The rapid sequence comparison algorithms we have developed also provide additional tools for evaluating the statistical significance of an alignment. There are approximately 5000 protein sequences, with 1.1 million amino acid residues, in the NBRF protein sequence library, and any computer program that searches the library by calculating a similarity score for each sequence in the library will find a highest scoring sequence, regardless of whether the alignment between the query and library sequence is biologically meaningful or not. Accompanying the previous version of FASTP was a program for the evaluation of statistical significance, RDF, which compares one sequence with randomly permuted versions of the potentially related sequence.

We have written a new version of RDF (RDF2) that has several improvements. (i) RDF2 calculates three scores for each shuffled sequence: one from the best single initial region (as found by FASTP), a second from the joined initial regions (used by FASTA), and a third from the optimized diagonal. (ii) RDF2 can be used to evaluate amino acid or DNA sequences and allows the user to specify the scoring matrix to be employed. Thus sequences found using the PAM250 scoring matrix can be evaluated using the identity or genetic code matrix. (iii) The user may specify either a global or local shuffle routine.

Locally biased amino acid or nucleotide composition is perhaps the most common reason for high similarity scores of dubious biological significance (10). High scoring alignments between query and library sequences may be due to patches of hydrophobic or charged amino acid residues or to A + T- or G + C-rich regions in DNA. A simple Monte Carlo shuffle analysis that constructs random sequences by taking each residue in one sequence and placing it randomly along the length of the new sequence will break up these patches of biased composition. As a result, the scores of the shuffled sequences may be much lower than those of the unshuffled sequence, and the sequences will appear to be related. Alternatively, shuffled sequences can be constructed by permuting small blocks of 10 or 20 residues so that, while the order of the sequence is destroyed, the local composition is not. By shuffling the residues within short blocks along the sequence, patches of G + C- or A + T-rich regions in DNA, for example, are undisturbed. Evaluating significance with a local shuffle is more stringent than the global approach, and there may be some circumstances in which both should be used in conjunction. Whereas two proteins that share a common evolutionary ancestor may have clearly significant similarity scores using either shuffling strategy, proteins related because of secondary structure or hydrophobic profile may have similarity scores whose significance decreases dramatically when the results of global and local shuffling are compared.

Implementation. The FASTA/LFASTA package of sequence analysis tools is written in the C programming language and has been implemented under the Unix, VAX/VMS, and IBM PC DOS operating systems. Versions of the program that run on the IBM PC are limited to query se-

Table 1. FASTA and FASTP initial scores of the T-cell receptor (RWMSAV) versus the NBRF data base

NBRF code	Sequence	Initial score	
		FASTA	FASTP
RWHUAV	T-cell receptor α chain	155	98
K1HURE	Ig κ chain V-I region	127	111
KVMS50	Ig κ chain V region	149	62
KVMSM6	Ig κ chain precursor V regions	141	64
KVRB29	Ig κ chain V region	126	54
L3HUSH	Ig λ chain V-III region	90	47
KVMS41	Ig κ chain precursor V region	87	87
RWMSBV	T-cell receptor β -chain precursor	94	94
RWHUVY	T-cell receptor β -chain precursor	91	59
RWHUGV	T-cell receptor γ -chain precursor	87	61
RWHUT4	T-cell surface glycoprotein T4	86	63
RWMSVB	T-cell receptor γ -chain precursor	71	41
HVMS44	Ig heavy-chain V region	67	36
G1HUDW	Ig heavy-chain V-II region	62	35

The average FASTP score = 26.1 ± 6.8 (mean \pm SD). The average FASTA score = 26.2 ± 7.2 (mean \pm SD). The mean and SD were computed excluding scores >54 . V, Variable.

quences of 2000 residues; library sequences can be any length. Copies of the program are available from the authors.

Although FASTA and LFASTA were designed for protein and DNA sequence comparison, they use a general method that can be applied to any alphabet with arbitrary match/mismatch scoring values. All the scoring parameters, including match/mismatch values, values for the first residue in a gap and subsequent residues in the gap, and other parameters that control the number of sequences to be saved and the histogram intervals, can be specified without changing the program.

EXAMPLES

Comparison of FASTA with FASTP. To demonstrate the superiority of the FASTA method for computing the initial score, we compared the protein sequence of a T-cell receptor α chain (NBRF code RWMSAV) with all sequences in the NBRF protein data base[†] and computed initial scores with both the present and previous methods. The T-cell receptor is a member of the immunoglobulin superfamily; in Release 12.0 of the data base, this superfamily has 203 members. FASTP placed 160 immunoglobulin superfamily sequences in the 200 top-scoring sequences; 57 related sequences received initial scores less than four standard deviations above the mean score. FASTA placed 180 superfamily members in the 200 top-scoring sequences; only 20 related sequences scored below four standard deviations above the mean. Table 1 contains specific examples from this data base search. Although there is often little difference in the two methods, this example shows that in a number of cases the new method obtains significantly higher scores between related sequences.

Nucleic Acid Data Base Search. FASTA can also be used to search DNA sequence data bases, either by comparing a DNA query sequence to the DNA library or by comparing an amino acid query sequence to the DNA library by translating each library DNA sequence in all six possible reading frames. We compared the 660-nucleotide rat transforming growth factor type α mRNA (GenBank locus RATTGFA) with all the mammalian sequences in Release 48 of GenBank[‡]. We set $ktup = 4$ (see *Methods*), and the search was completed in under 15 min on an IBM PCAT microcom-

Table 2. DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank locus	Sequence	Score	
		Initial	Optimized
HUMTGFAM	Human TGF mRNA	1336	1618
HUMTGFA2	Human TGF gene (exon 2)	354	366
HUMTGFA1	Human TGF gene (5' end)	224	381
MUSRGEB3	Mouse 18S-5.8S-28S rRNA gene	140	107
MUSRGE52	Mouse 18S-5.8S-28S rRNA gene	140	107
MUSMHDD	MHC class I H-2D	122	78
HUMMETIF1	Metallothionein (MT) _I gene	116	92
MUSRGLP	45S rRNA (5' end)	115	83
HUMPS2	pS2 mRNA	105	106
MUSC1A11	α -1 type I procollagen	86	89

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; MHC, major histocompatibility complex.

puter. The 10 top-scoring library sequences are shown in Table 2. Although it can be seen that the 3 top-scoring sequences are clearly related to RATTGFA, there are other high-scoring sequences that are probably not related, and the mouse epidermal growth factor, found in the translated data base search (Table 3), is not found among the top-scoring sequences.

To further examine the similarity detected between RATTGFA and MUSRGEB3, a mouse rRNA gene cluster, we used the RDF2 program for Monte Carlo analysis of statistical significance (the window for local shuffling was set to 10 bases). Of the 50 shuffled comparisons (data not shown), 1 obtained an initial score greater than 140 (the observed initial score), and 9 shuffled sequences obtained optimized scores greater than 107 (the observed optimized score). Therefore, the similarity between RATTGFA and MUSRGEB3 is unlikely to be significant.

Translated Nucleic Acid Data Base Search. When searching for sequences that encode proteins, amino acid sequence comparisons are substantially more sensitive than DNA sequence comparisons because one can use scoring matrices like the PAM250 matrix that discriminate between conservative and nonconservative substitutions. A variant of FASTA, TFASTA, can be used to compare a protein sequence to a DNA sequence library; it translates the DNA sequences into each of six possible reading frames "on-the-fly." TFASTA translates the DNA sequences from beginning to end; it includes both intron and exon sequences in the translated protein sequence; termination codons are translated into unknown (X) amino acids. Table 3 shows the results of a translating search of the mammalian sequences in the GenBank DNA data base using the RATTGFA protein sequence as the query and $ktup = 1$. In the translated search, the mouse epidermal growth factor now obtains an initial score higher than any unrelated sequences; however, HUMTGFA1, which was found in the DNA data base search but only contains 13 translated codons, is no longer among the top scoring sequences.

Local Similarities. Fig. 2 displays the output of a local similarity analysis ($ktup = 4$) of CHPHBA1M, a chimpanzee α 1-globin mRNA, and RABHBAPT, a rabbit α -globin gene, including the complete coding sequence and a flanking pseudo- θ -globin gene. LFASTA can either display a graphic matrix style plot of the local homologies (Fig. 2A) or the alignments themselves (Fig. 2B). The right-most three alignments (Fig. 2A) match the corresponding regions of the mRNA to exon subsequences from the pseudogene. We note that the FASTA initial score for the comparison of CHPH-

[†]Protein Identification Resource (1987) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 12.

[‡]EMBL/GenBank Genetic Sequence Database (1987) (Intelligenetics, Mountain View, CA), Tape Release 48.

Table 3. Translated DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank locus	Sequence	Frame	Score	
			Initial	Optimized
RATTGFA	Rat TGF type α	1	816	816
HUMTGFA	Human TGF mRNA	2	671	770
HUMTGFA2	Human TGF gene	1	204	205
MUSEGF	Mouse EGF mRNA	3	93	129
MUSMHA3	Mouse MHC class II H2-1A _B	1	91	58
MUSIGCD17	Mouse Ig germ-line DJC region	3'	85	48
HUMESTR	Human estrogen receptor	3	83	65
RATINSI	Rat insulin 1 (<i>Ins-1</i>) gene	2	81	63
MUSTHYS1	Mouse thymidylate synthase	2	80	63
HUMPNU3	Human purine nucleoside phosphorylase	1'	80	52

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; EGF, epidermal growth factor; D, diversity; J, joining; C, constant; MHC, major histocompatibility complex.

BA1M and RABHBAPT would be based on the three globin gene exons, while the FASTP initial score would be based on a single conserved exon.

The Smith-Waterman optimization used in the LFASTA program allows the detection of more subtle features than can be detected by the eye using a graphic matrix plot, because the path traced is locally optimal, even though it may only have a slightly higher density of identities and conservative replacements. Fig. 3 shows a plot from a local similarity self-comparison of the myosin heavy chain from the nematode *Caenorhabditis elegans* (MWKW) using the PAM250 matrix. The amino-terminal half of the molecule forms a large globular head without any periodic structure; the solid line down the main diagonal represents the expected identity of the sequence with itself. The symmetrical parallel lines along the carboxyl-terminal half of the molecule correspond to the 28-residue repeat responsible for the α -helical coiled-coil structure of the rod segment.

DISCUSSION

In searching a data base, one is attempting to measure relatedness; in aligning two homologous sequences, one is

trying to choose the most likely set of mutations since their divergence from a common ancestral sequence. Thus any tool for the analysis of sequence similarities must contain within it an implicit model of molecular evolution. An algorithm that guarantees the optimality of its alignments based on a set of scoring rules must be judged on how well these rules fit our current understanding of the process of molecular evolution. Algorithms that sacrifice realism to achieve greater efficiency, regardless of their mathematical rigor, require careful empirical evaluation.

Even though the tools we have developed use rigorous algorithms at each step and incorporate a realistic model of evolution, their hierarchical nature make them heuristic. The original FASTP program has had the benefit of extensive use and evaluation by a wide variety of scientists. The FASTA program exploits refinements of the previous approach that result in a significant improvement in sensitivity. The LFASTA local similarity analysis program is also a logical extension of the FASTP approach.

Because of the trade-offs between sensitivity and selectivity in data base searches, the results of any search, and particularly those that result in alignment scores that are not clearly separated from the distribution of all library sequence

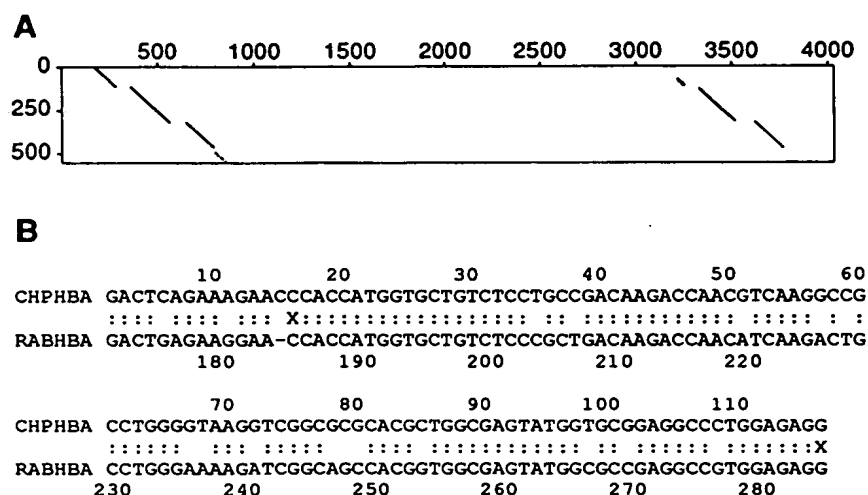


FIG. 2. Local comparison of an α -globin mRNA sequence with an α -globin gene cluster. An ape α_1 -globin mRNA sequence (GenBank sequence CHPHBA1M) was compared with a rabbit α -globin gene sequence (RABHBAPT) containing a second pseudo- θ -globin gene using the LFASTA program. (A) A plot of the homologous regions shared by the two sequences. (B) One of the alignments between the mRNA sequence and the rabbit α -globin gene (nucleotides 171–855). Three other alignments between the mRNA sequence and the α -globin gene and three alignments between the pseudo- θ -globin gene (nucleotides 3200–3770) were calculated but are not shown. There is 84.3% identity in the 115 nucleotide overlap. The initial region and optimized scores using LFASTA are 284 and 304, respectively. X denotes the ends of the initial region found by LFASTA.

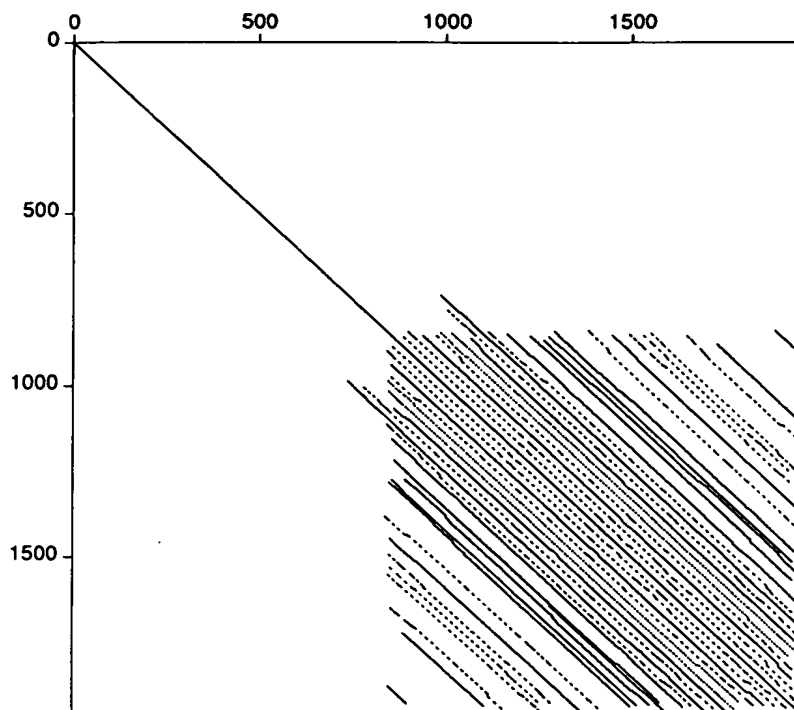


FIG. 3. Repeated structure in the myosin heavy chain. LFASTA was used to compare the *Caenorhabditis elegans* myosin heavy chain protein sequence (NBRF code MWKW) with itself using the PAM250 scoring matrix. The solid, dashed, and dotted lines denote decreasing similarity scores. The solid lines had initial region scores greater than 80 and optimized local scores greater than 150; the longer dashed lines had initial region and optimized local scores greater than 65 and 120, respectively, and the shorter dashed lines had initial region and optimized local scores greater than 50 and 100, respectively. Homologous regions with lower scores are plotted with dots.

scores, must be carefully evaluated (1, 11). The Monte Carlo analysis of statistical significance provided by a program such as RDF2 can often be critical in evaluating a borderline similarity. Previously we suggested ranges of z values [(observed score - mean of shuffled scores)/standard deviation of shuffled scores] corresponding to approximate significance levels. However the z values determined in a Monte Carlo analysis become less useful as the distribution of shuffled scores diverges from a normal distribution, as is found with FASTA. Therefore, we now focus on the highest scores of the shuffled sequences. For example, if in 50 shuffled comparisons, several random scores are as high or higher than the observed score, then the observed similarity is not a particularly unlikely event. One can have more confidence if in 200 shuffled comparisons, no random score approaches the observed score. In general, our experience has led us to be conservative in evaluating an observed similarity in an unlikely biological context.

These programs provide a group of sequence analysis tools that use a consistent measure for scoring similarity and constructing alignments. FASTA, RDF2, and LFASTA all use the same scoring matrices and similar alignment algorithms, so that potentially related library sequences discov-

ered after the search of a sequence data base can be evaluated further from a variety of perspectives. In addition, LFASTA can also show alternative alignments between sequences with periodic structures or duplications.

1. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435-1441.
2. Dumas, J. P. & Ninio, J. (1982) *Nucleic Acids Res.* **10**, 197-206.
3. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726-730.
4. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345-352.
5. Needleman, S. & Wunsch, C. (1970) *J. Mol. Biol.* **48**, 444-453.
6. Smith, T. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195-197.
7. Maizel, J. & Lenk, R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7665-7669.
8. Goad, W. & Kanehisa, M. (1982) *Nucleic Acids Res.* **10**, 247-263.
9. Sellers, P. H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3041.
10. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215-226.
11. Doolittle, R. (1981) *Science* **214**, 149-159.

Exhibit 29

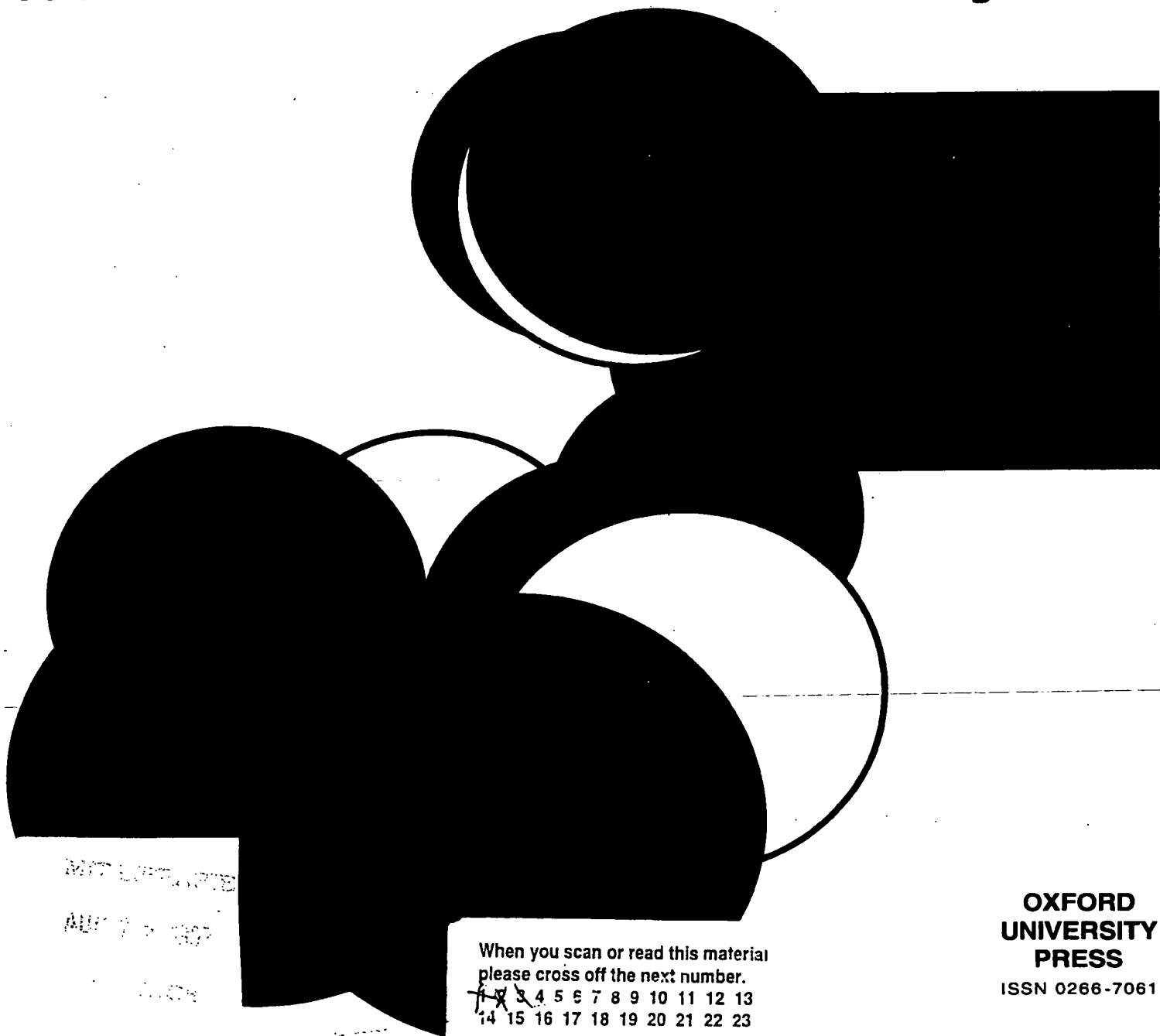
CABIOS

64

COMPUTER
APPLICATIONS
IN THE
BIOSCIENCES

Volume 13 Number 4

August 1997



NOT RECORDED

AUG 11 1997

11/1/97

When you scan or read this material
please cross off the next number.

~~1~~ 2 3 4 5 6 7 8 9 10 11 12 13
14 15 16 17 18 19 20 21 22 23

OXFORD
UNIVERSITY
PRESS

ISSN 0266-7061

Identifying distantly related protein sequences

William R. Pearson

Introduction

The most powerful method available today for inferring the biological function of a gene (or the protein that it encodes) from its sequence is similarity searching on protein and DNA sequence databases. With the development of rapid methods for sequence comparison, both with heuristic algorithms and powerful parallel computers, discoveries based solely on sequence homology have become routine. Indeed, the vast majority of the gene identifications in the recent descriptions of the *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Mycoplasma genitalium* (Fraser *et al.*, 1995), yeast (Dujon, 1996) and *Methanococcus janesscii* (Bult *et al.*, 1996) genomes are based only on protein sequence similarity. As more complete genomes become available, protein sequence comparison will become an even more powerful tool for understanding biological function.

Protein sequence comparison is a powerful tool because of the enormous amount of information that is preserved throughout the evolutionary process. For many protein sequences, an evolutionary history can be traced back 1–2.5 billion years. Proteins that share a common ancestor are called homologous. Sequence comparison is most informative when it detects homologous proteins. Homologous proteins always share a common three-dimensional folding structure and they often share common active sites or binding domains. Frequently, homologous proteins share common functions, but sometimes they do not. Our ability to characterize the biological properties of a protein based on sequence data alone stems almost exclusively from properties conserved through evolutionary time. Predictions of common properties for non-homologous proteins—similarities that have arisen by convergence—are much less reliable.

While sequence similarity searching is a routine method for characterizing newly determined DNA and protein sequences, researchers sometimes fail to exploit fully the information that is available from similarity searches of protein sequence databases. This review examines two strategies for using similarity search information more effectively: (i) looking for alignments that span an entire folding domain, rather than a short sequence motif, and (ii)

re-examining sequences with high, but not statistically significant, similarity scores. For a broader perspective on sequence comparison and identification of homologous proteins, see Altschul *et al.* (1994) and Pearson (1996).

Members of the trypsin-like serine protease superfamily ('trypsin-like' distinguishes these serine proteases from other serine protease families—notably the subtilisins—that use serine in the active site but have very different structures and thus are not homologous) provide a classic example of a family of proteins with a highly conserved active site. While highly conserved motifs from this site are informative, serine proteases share similarity throughout the length of the protease domain, not just around the active site residues.

The trypsin-like serine protease family is quite diverse, with a number of very distantly related homologues. Thus, it can be difficult to demonstrate that *Streptomyces griseus* protease A and protease B are homologous based on sequence similarity alone. The second part of this review shows that by carefully re-examining sequences with high-scoring, but not statistically significant, similarity scores, it is possible to identify several proteins that share significant similarity with both the mammalian trypsin-like serine proteases and their distant prokaryotic homologues.

Motifs, homology, and the serine proteases

A common misconception in protein sequence comparison is that homologous proteins share sequence similarity mostly (or only) near the active site regions or other functional domains in a protein. This partly accounts for the popularity of databases of sequence motifs, such as PROSITE (Bairoch, 1991), which tabulate amino acid patterns that can be used to identify most of the members of a protein family. For features that result from convergence to a common property, such as glycosylation and phosphorylation sites, sequence motifs are uniquely informative. However, for features that result from divergence from a common ancestor, such as the serine protease active site residues, sequence motifs provide only a highly abstracted summary of the sequence conservation in a family. Because they share a common three-dimensional structure, homologous proteins share sequence similarity over large regions—typically the entire protein fold.

The trypsin-like serine protease superfamily is a classic example of a protein family whose members share several simple motifs that are diagnostic for the family (Figure 1).

Department of Biochemistry, Jordan Hall #440, University of Virginia,
Charlottesville, VA 22908, USA
E-mail: wrp@virginia.EDU

```

ID  TRYPSIN_HIS; PATTERN.
AC  PS00134;
DE  Serine proteases, trypsin family, histidine active site.
PA  [LIVM]-[ST]-A-[STAG]-H-C.
NR  /TOTAL=158(158); /POSITIVE=154(154); /UNKNOWN=2(2); /FALSE_POS=2(2);
NR  /FALSE_NEG=11(11);
CC  /TAXO-RANGE=??EP?; /MAX-REPEAT=1;
CC  /SITE=5,active_site;

ID  TRYPSIN_SER; PATTERN.
AC  PS00135;
DE  Serine proteases, trypsin family, serine active site.
PA  G-D-S-G-G.
NR  /TOTAL=160(160); /POSITIVE=151(151); /UNKNOWN=1(1); /FALSE_POS=8(8);
NR  /FALSE_NEG=16(16);
CC  /TAXO-RANGE=??EP?; /MAX-REPEAT=1;
CC  /SITE=3,active_site;

```

Fig. 1. Patterns for serine proteases. Patterns from PROSITE that identify 152/163 TRYPSIN_HIS or 143/159 TRYPSIN_SER members of the trypsin-like serine protease protein family.

Serine proteases cleave peptide bonds using a 'catalytic triad' of histidine, serine and aspartic acid that are required for the protease function. Because these residues are so highly conserved, patterns that focus on two of the regions (Figure 1) can be used to identify every member of the serine protease family. (The subtilisin-like serine proteases use exactly the same catalytic triad, but the families are non-homologous with very different three-dimensional structures.)

Most members of the trypsin-like serine protease superfamily are readily identified by sequence similarity searching. The results from a typical protein database search using the Smith-Waterman algorithm (Smith and Waterman, 1981) are shown in Figure 2. All of the eukaryotic trypsin-like serine proteases share statistically significant similarity with the bovine trypsin query sequence. However, as is often the case for divergent protein families, some prokaryotic members of the family do not share statistically significant similarity with bovine trypsin. These sequences are italicized in Figure 2; their membership in the serine protease family is usually inferred from their common three-dimensional structures (Figure 5).

The absolute conservation of residues in the 'catalytic triad' might suggest that sequence similarities shared by members of this family are limited to those regions. Indeed, two of the four 'High-Scoring segment Pairs' (Altschul *et al.*, 1994) reported by BLASTP correspond to TRYP_HIS and TRYP_SER regions (Figure 3). However, similarity in the serine proteases extends from one end of the protein to the other, with conservation throughout the sequence. Indeed, many parts of protein are conserved more strongly than the region around the aspartic acid in the catalytic triad (Figure 3). Thus, while the residues in the catalytic triad are an essential feature for a functional serine protease, it is the

serine protease fold (two domains containing anti-parallel β barrels; Figure 5) that is required to bring these residues together. The evolutionary pressure to conserve the trypsin-like serine protease fold ensures that the folding domains share similar amino acids.

The requirement for a common folded structure in homologous proteins usually causes similarities to extend from one end of the protein to the other. With the exception of mosaic proteins that are the result of recent exon shuffling (Doolittle, 1995), optimal local sequence similarity is rarely confined only to a portion of two homologous sequences. (In mosaic proteins, the similarity extends throughout the exon-shuffled domain.) In general, it is incorrect to speak of homology at the N terminus or C terminus, even though only a portion of the protein may be aligned in 'High Scoring segment Pairs' by BLASTP. Indeed, the length of the locally similar region can sometimes be used to distinguish low-scoring related sequences from high-scoring unrelated sequences. Thus, all but two of the library sequences (including four with expectation values >0.02) that align over $>80\%$ of the length of the TRYP_BOVIN query sequence are members of the trypsin-like serine protease family. Figure 4 displays the locally similar regions for the related and unrelated sequences in Figure 2; the highest scoring unrelated sequences tend to have relatively short (<100 residue) regions of higher similarity ($\sim 30\%$ identical), while related sequences have longer (140–300 residue) aligned regions, sometimes with lower (25%) sequence identity. In general, alignments with longer, lower identity are more significant than those with shorter, higher identity.

The requirement for similarity over a large region is more evident when three-dimensional structures are examined. TRYP_BOVIN (structure not shown), TRYP_STRGR

LOCUS	Description	len	score	E(51,780)
TRYP_BOVIN	trypsinogen (EC 3.4.21.4).	229	1559	0
TRY2_HUMAN	trypsinogen II	247	1201	0
TRYP_PLEPL	trypsin	250	788	0
KLK2_HUMAN	glandular kallikrein 2	261	665	0
RVVA_VIPRU	vipera russelli proteinase	236	637	0
TRY1_ANOGA	trypsin 1	274	600	10 ⁻³²
TRYA_DROME	trypsin alpha	256	579	10 ⁻³¹
FA9_RAT	coagulation factor IX	282	573	10 ⁻³⁰
PLMN_PIG	plasminogen	790	569	10 ⁻³⁰
TRY5_ANOGA	trypsin 5	274	550	10 ⁻²⁹
TRYP_FUSOX	trypsin	248	541	10 ⁻²⁸
FA7_RABIT	coagulation factor VII	443	519	10 ⁻²⁷
URTB_DESRO	salivary plasminogen activator β	431	508	10 ⁻²⁶
ACRO_PIG	acrosin	415	501	10 ⁻²⁶
PRTC_HUMAN	protein C	461	494	10 ⁻²⁵
TRYM_CANFA	mastocytoma protease	269	484	10 ⁻²⁵
TRYP_STRGR	trypsin	259	410	10 ⁻²⁰
HGF_HUMAN	hepatocyte growth factor prec.	728	397	10 ⁻¹⁸
ACH1_LONAC	achelase I protease	213	352	10 ⁻¹⁶
CERC_SCHMA	cercarial protease	264	203	10 ⁻⁶
CO2_HUMAN	complement C2	752	198	10 ⁻⁵
CFAB_MOUSE	complement factor B	761	170	0.00041
PRTZ_BOVIN	vitamin K-dependent protein Z	396	142	0.015
LORI_MOUSE	loricrin.	481	125	0.24
GSEP_BACLI	glutamyl endopeptidase	316	118	0.45
KRUC_SHEEP	keratin, ultra high-sulfur matrix	182	107	1.3
PRLA_LYSEN	alpha-lytic protease	397	107	3.1
AGI_URTDI	lectin/endochitinase precursor	372	105	3.9
KCR8_YEAST	prob. serine/threonine-protein kin.	603	107	4.7
G156_PARPR	156g surface protein precursor	715	117	5.0
YLK3_CAEEL	putative ser./thr.-protein kinase	895	114	5.4
AMY_CLOAB	putative alpha-amylase	469	104	5.7
AGI_HORVU	root-specific lectin precursor	212	98	6.2
YB9X_YEAST	hypothetical trp-asp repeats	878	105	9.5
PRTS_MOUSE	vitamin k-dependent protein S	675	103	9.8
DLK_HUMAN	delta-like protein	383	99	9.9
PRTB_STRGR	streptogrisin B (S. gris. prot. A)	299	94	16.
PRTA_STRGR	streptogrisin A (S. gris. prot. A)	297	85	64.

Fig. 2. Serine protease search—high-scoring sequences. High-scoring sequences from a search of SwissProt (Bairoch and Boeckmann 1991; release 33, April 1996) with TRYP_BOVIN. Only 10% of the database sequences with $E() < 10^{-6}$ are shown. Trypsin-like serine proteases with $E() > 0.02$ are in italics.

(Figure 5, 1sbt) and PRTA_STRGR (1sgc) share a very similar all- β fold with symmetrical β barrel structures and two short α helices. Very little of this structure is directly involved in forming the catalytic triad in the active site; yet the entire fold is conserved, thus requiring conservation of an amino acid sequence that adopts this fold.

Although almost all vertebrate trypsin-like serine proteases share significant sequence similarity with bovine trypsin, most bacterial serine proteases do not. For example, the similarity score for alignment of bovine trypsin with *S. griseus* protease A is not statistically significant ($E() < 64$), even though the structures of the two enzymes are very similar (Figure 5). Thus, while statistically significant similarity generally implies common ancestry, and thus common three-dimensional structure [the most common exceptions to this rule are regions with very low amino acid complexity, e.g. YSGGGSSCGGGYSGGGSSCGGGSSGGG from LORI_MOUSE (Altschul *et al.*, 1994)], lack of statistically significant similarity does not imply non-homology.

Figure 5 also shows the structures of two non-homologous

proteins. Subtilisin (1sbt) is included because it is an example of 'convergent' evolution (Doolittle, 1994); subtilisin uses the same triad of catalytic residues (Asp, His and Ser) to cleave peptide bonds, but shares no structural similarity beyond the geometry of the active site of the enzyme. Subtilisin and subtilisin-like serine proteases are not homologous to the trypsin-like serine proteases. As expected, the different structures share no statistically significant sequence similarity (1500 random sequences from SwissProt would be expected to have a better similarity score than that obtained in the trypsin/subtilisin comparison).

Likewise, high-scoring sequences that are not homologous to trypsin-like serine proteases rarely share structural similarity to the family, despite their 'strong' similarity. Wheat germ agglutinin (7wga) is the most similar non-serine protease sequence in the NRL_3D database of sequences whose structures are known, yet it does not contain a single β sheet. With the exception of membrane-spanning proteins, which frequently share hydrophobic regions with other unrelated membrane proteins, high sequence similarity—in

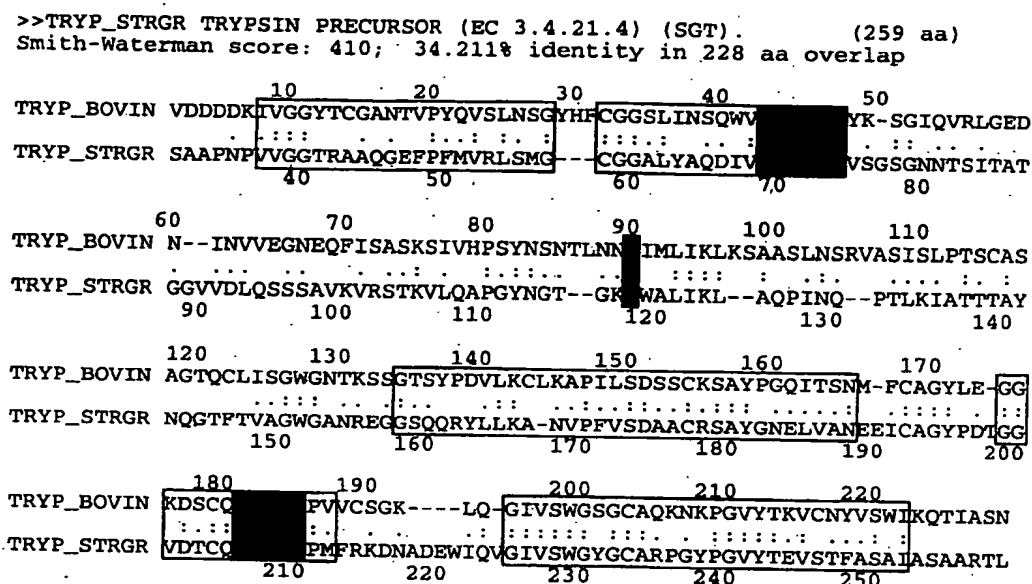


Fig. 3. Alignment of serine proteases. Alignment of bovine trypsinogen (TRYP_BOVIN) and *S. griseus* trypsin (TRYP_STRGR). Shaded boxes indicate the TRYP_HIS and TRYP_SER patterns shown in Figure 1 and the conserved 'D' that is the third component of the catalytic triad. Unshaded boxes indicate the consistent 'High Scoring segment Pairs' reported by BLASTP.

the absence of homology—provides no information about structural similarity.

Using statistical significance to explore distant relationships

A major advance in sequence identification by similarity searching has been the development of accurate statistical estimates for similarity scores (Altschul *et al.*, 1994). Since the similarity score from comparison of TRYP_BOVIN and TRYP_STRGR has an expectation value of $E() < 10^{-20}$, we conclude that these two sequences share similarity that would never be obtained by chance (or obtained once in 10^{20} searches of a database the size of SwissProt), and thus their similarity reflects a common ancestry for the two sequences. Current versions of the FASTA package of sequence comparison programs (version 2 and 3) include accurate statistical estimates for both FASTA and SSEARCH (Smith-Waterman) similarity scores (Pearson, 1996). Careful analysis of the high-scoring non-homologous sequences can be used both to confirm that the statistical estimates are reliable and to explore distantly related members of a protein family.

Identifying the highest-scoring non-homologous sequences in a database search may seem difficult if the protein family is very diverse. However, additional searches with high-scoring, but possibly unrelated sequences can be used to separate high-scoring unrelated sequences from distantly related sequences. Additional searches with high-scoring

unrelated sequences will typically produce 'matches' with unrelated sequences, while additional searches with distantly-related sequences will produce 'matches' to protein family members. If the statistical estimates are accurate, high-scoring unrelated sequences will have $E()$ values of ~ 1.0 , since one highest scoring sequence is expected in every search. If the $E()$ value for the highest scoring unrelated sequences are unexpectedly low and the sequences do not contain low-complexity simple sequence repeats, additional searches can be carried out with higher gap penalties.

Bovine trypsin (TRYP_BOVIN) shares statistically significant similarity with every full-length mammalian serine protease, but the bacterial alpha-lytic protease (PSLA_LYSEN) or *S. griseus* protease A or protease B do not share significant similarity with bovine trypsin. There is no question that these proteins are homologous to the mammalian trypsin-like enzymes because of their strong structural similarity (Figure 5). However, in the absence of high-resolution structural data, how can one decide whether a high-scoring, but not significantly similar, sequence is homologous?

Additional searches with the highest scoring, non-significant matches allow us to identify additional members of the family. A search with PRTZ_BOVIN, which has a marginally significant score, shows strong similarity ($E()$ values $< 10^{-10}$) with a variety of other members of the family, thus confirming its homology. LORI_MOUSE gives a different result; while many serine proteases are highly ranked with

LOCUS E() % ident.

TRYP_BOVIN	0	100.0
TRY2_HUMAN	0	75.0
TRYP_PLEPL	0	45.7
KLK2_HUMAN	0	43.5
RVVA_VIPRU	0	40.9
TRY1_ANOGA	10 ⁻³²	39.9
TRYA_DROME	10 ⁻³¹	42.1
FA9_RAT	10 ⁻³⁰	40.9
PLMN_PIG	10 ⁻³⁰	40.8
TRY5_ANOGA	10 ⁻²⁸	38.7
TRYP_FUSOX	10 ⁻²⁸	41.6
FA7_RABIT	10 ⁻²⁷	37.2
URTB_DESRO	10 ⁻²⁷	38.2
ACRO_PIG	10 ⁻²⁶	35.7
PRTC_HUMAN	10 ⁻²⁶	34.5
TRYM_CANFA	10 ⁻²⁵	37.5
TRYP_STRGR	10 ⁻²⁰	34.2
HGF_HUMAN	10 ⁻¹⁸	31.6
ACH1_LONAC	10 ⁻¹⁶	33.5
CERC_SCHMA	10 ⁻⁶	26.9
CO2_HUMAN	10 ⁻⁵	26.1
CFAB_MOUSE	10 ⁻³	24.0
PRTZ_BOVIN	0.015	25.2
LORI_MOUSE	0.24	33.7
GSEP_BACLI	0.45	20.6
KRUC_SHEEP	1.3	27.9
PRLA_LYSEN	3.1	21.5
AGI_URTDI	3.9	26.1
KCR8_YEAST	4.7	33.3
G156_PARPR	5.0	31.2
YLK3_CAEEL	5.4	25.9
AMY_CLOAB	5.7	23.3
AGI_HORVU	6.2	24.8
YB9X_YEAST	9.5	32.3
PRTS_MOUSE	9.8	28.4
DLK_HUMAN	9.9	34.2
PRTB_STRGR	16.	24.0
PRTA_STRGR	64.	23.4

Fig. 4. Serine protease alignments The alignments of each of the high-scoring sequences reported in Figure 2 are indicated by mapping back to the TRYP_BOVIN query sequence. Thus, alignment of TRYP_BOVIN with itself extends from the beginning to the end of the query sequence; alignment of TRYP_BOVIN and TRYA_DROME extends over 85% of the TRYP_BOVIN query sequence. Members of the family with E() > 0.02 are italicized. The E() value and percent identity are also shown. The ssearch -m 4 option was used to produce this figure.

significant similarity, the sequence alignments contain a repeated glycine and serine motif. Thus, LORI_MOUSE is not homologous; it contains an unusual simple amino acid repeat sequence. On the other hand, GSEP_BACLI shares strong similarity with several bacterial serine proteases ($E() < 10^{-9}$) and weaker, but significant similarity with TRYP_SACER and TRYP_FUSOX, *Streptomyces* and yeast trypsins with very strong similarity to bovine trypsin. GSEP_BACLI is, therefore, a member of the trypsin-like serine protease family.

A search with alpha-lytic protease reveals a second group of closely related serine proteases, which includes *S.griseus* protease A and protease B. While none of the sequences in

Figure 2 have significant similarity with PRLA_LYSEN, GLUP_STRGR, an *S.griseus* glutamyl endopeptidase, shares strong similarity with the *S.griseus* protease A and B, alpha-lytic protease, and weaker, but significant similarity with TRYA_DROME and several other *Drosophila* serine proteases (Figure 6). The insect sequences share strong similarity to mammalian trypsin-like serine proteases (Figure 2). Thus, by carefully exploring sequences with high, but not statistically significant, similarity scores, it is possible to construct statistically significant links between very distantly related serine proteases.

Distant sequence relationships can thus be established by moving from sequence A to significantly similar sequence B,

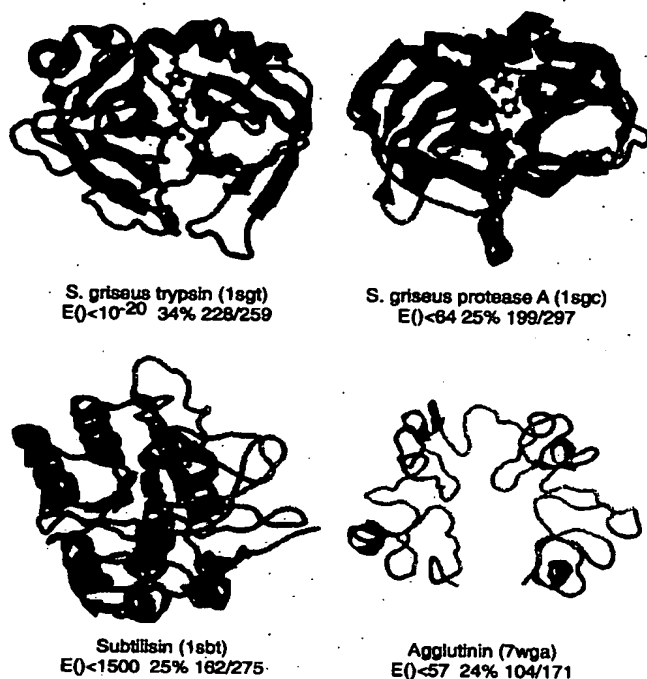


Fig. 5. Structures—homologous, convergent and unrelated. The structures of two members (1sgt, 1sgc) of the trypsin-like serine protease family are shown, along with subtilisin (1sbt)—a non-trypsin-like serine protease—and wheat germ agglutinin (7wga), one of the highest scoring non-serine proteases in the NRL_3D database (release 20) of sequences whose structures are known. Serine protease structures are aligned to present a similar view of the catalytic site. The expectation values shown are based on a comparison of bovine trypsin (TRYP_BOVIN) to the SwissProt (release 33) protein sequence database. Also shown are the percent identity and the length of the similar region with respect to the length of the sequence of the structure shown.

and then from B to C, even though A does not share significant similarity with C. The strategy is effective because of the implicit evolutionary tree that connects all the members of a protein family. Thus, in Figure 7, a sequence on a relatively short branch, TRYA_DROME, can be used to establish significant relationships with very diverse members of the family. For large and diverse protein families, it is usually easy to identify a number of 'less-divergent' family members that can be used to link distant branches of the tree. Naturally, such inferences are more reliable if statistically significant similarity scores are produced with different sets of scoring matrices and gap penalties, and if they are established with several different linking sequences.

A phylogenetic tree was produced from selected vertebrate, invertebrate and prokaryotic trypsin-like serine proteases. Sequences were aligned using ClustalW (Thompson *et al.*, 1994) and protein distances estimated and distance trees built using the PHYLIP package (Felsenstein, 1989). The three numbers to the right of the sequence names report the statistical significance of the alignment score between the sequence and bovine trypsin (TRYP_BOVIN), *Drosophila* trypsin A (TRYA_DROME) and *S.griseus* glutamyl endopeptidase (GLUP_STRGR), respectively. MPR_BACSU is an example of another sequence that links eukaryotic and prokaryotic serine proteases, although it does not share statistically significant similarity with the three query sequences used for expectation values here.

Summary

Protein sequence comparison is the most powerful tool available today for inferring structure and function from sequences because of the constraints of protein evolution—a

LOCUS	Description	len	score	E(51,934)
GLUP_STRGR	glutamyl endopeptidase II	188	1223	0
SFA1_STRFR	serine protease 1	357	1019	0
PRTA_STRGR	streptogrisin A	297	681	0
PRTB_STRGR	streptogrisin B	299	624	10 ⁻³⁰
SFA2_STRFR	serine protease 2	174	583	10 ⁻²⁸
PRLA_LYSEN	alpha-lytic protease	397	349	10 ⁻¹⁴
SP1_RARFA	serine protease I	525	297	10 ⁻¹⁰
TRYA_DROME	trypsin alpha	256	160	0.0031
LORI_HUMAN	loricrin.	316	157	0.0057
LORI_MOUSE	loricrin.	481	160	0.0058
TRYB_DROME	trypsin beta	253	152	0.009
AIDA_ECOLI	adhesin AIDA-I	1286	155	0.032
TRYD_DROME	trypsin delta	253	139	0.054
GSEP_BACLI	glutamyl endopeptidase	316	140	0.059
TRYG_DROME	trypsin gamma	253	138	0.061
TRYP_FUSOX	trypsin	248	135	0.091
APMU_PIG	apomucin mucin core protein	1150	144	0.13
SLAP_CAUCR	S-layer paracryst. surf. prot	1025	142	0.15
TRY4_LUCCU	trypsin alpha-4	255	130	0.19

Fig. 6. From glutamyl endopeptidase to TRYA_DROME.

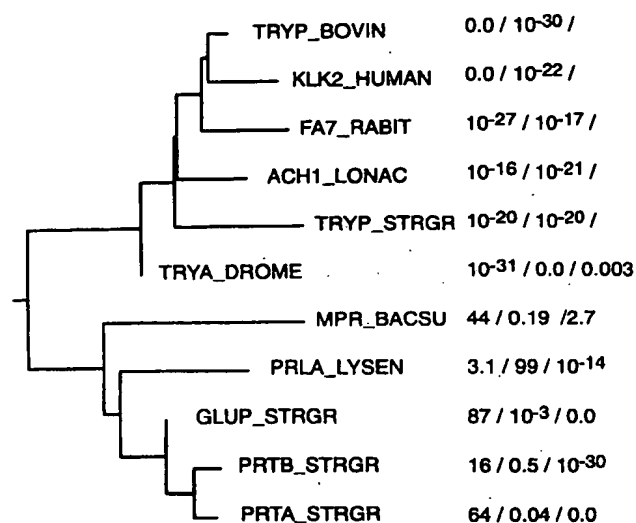


Fig. 7. Similarity and homology—a serine protease family tree.

protein must fold into a functional structure—which are reflected in its sequence. Protein sequence similarity can routinely be used to infer relationships between proteins that last shared a common ancestor 1–2.5 billion years ago. Our ability to identify distantly related proteins has improved over the past 5 years with the use of optimized scoring parameters (Pearson, 1995) and the development of accurate statistical estimates. In using sequence similarity to infer homology, one should remember the following.

1. Always compare protein sequences if the genes encode proteins. Protein sequence comparison will typically double the look-back time over DNA sequence comparison.
2. Homologous sequences are usually similar over an entire sequence or domain. Matches that are > 50% identical in a 20–40 amino acid region frequently occur by chance.
3. While most sequences that share statistically significant similarity ($E() < 0.02$) are homologous, many distantly related homologous sequences do not share significant homology. (Significant similarity in low-complexity regions does not imply homology.)
4. By focusing on the statistical significance of a similarity and identifying the highest scoring unrelated sequence in a database search, you can both confirm that the statistical estimates are accurate and potentially identify distantly related family members.
5. Homologous sequences share a common ancestor, and thus a common protein structure. Depending on the evolutionary distance and divergence path, two or more homologous sequences may have very few absolutely conserved residues. However, if homology has been inferred between A and B, between B and C, and between C and D, A and D must be homologous, even if they share no significant similarity when

compared directly. In evaluating the results of a similarity search, remember that there is an evolutionary tree that connects the family members.

Motifs revisited

This review argues that sequence similarity searching, rather than motif identification, is the most reliable method for identifying distantly related protein sequences. However, motif searches are frequently used to characterize a newly determined sequence. While motifs can be very valuable for identifying functional sites in a protein, one must be very careful in basing sequence identifications on motif patterns alone. Thus, if a newly determined protein sequence contains the G-D-S-G-G motif, but does not share strong similarity ($E() < 20$) with any of the hundreds of trypsin-like serine proteases in the protein databases, is it likely to be homologous to trypsin and share the same protein fold? It seems unlikely, since so many very distantly related members of the family are known. However, if a protein sequence shares high, but not significant ($0.02 < E() < 20$) sequence similarity with several distantly related members of the family, the presence of the two motifs in Figure 1 would provide strong supporting evidence that a new branch in the serine protease family had been found.

Alternatively, if a sequence shares significant similarity with proteins from several branches of the serine protease family tree, but does not contain the G-D-S-G-G motif, it is very likely that it adopts the serine protease protein fold, although it may not function as a protease. Thus, when enzymatic mechanisms are known, motifs can be used to confirm functional aspects of homologous proteins. However, in the absence of strong similarity to any member of a large protein family, motifs are unreliable for inferring protein homology.

References

- Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genet.*, 6, 119–129.
- Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, 19(Suppl.), 2241–2245.
- Bairoch,A. and Boeckmann,B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, 19(Suppl.), 2247–2249.
- Bult,C.J. et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273, 1058–1073.
- Doolittle,R.F. (1994) Convergent evolution: the need to be explicit. *Trends Biochem. Sci.*, 19, 15–18.
- Doolittle,R.F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem.*, 64, 287–314.
- Dujon,B. (1996) The Yeast Genome Project, what did we learn? *Trends Genet.*, 12, 263–270.
- Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, 5, 164–166.
- Fleischmann,R.D. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269, 496–512.
- Fraser,C.M. et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270, 397–403.
- Pearson,W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, 4, 1145–1160.

- Pearson, W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.*, **266**, 227–258.
- Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) ClustalW: Improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Received on November 21, 1996; revised on January 10, 1997; accepted on January 28, 1997

Exhibit 30

REVIEW

Structural basis of substrate specificity in the serine proteases

JOHN J. PERONA¹ AND CHARLES S. CRAIK

Departments of Pharmaceutical Chemistry and Biochemistry & Biophysics,
University of California, San Francisco, California 94143-0446

(RECEIVED July 25, 1994; ACCEPTED December 28, 1994)

Abstract

Structure-based mutational analysis of serine protease specificity has produced a large database of information useful in addressing biological function and in establishing a basis for targeted design efforts. Critical issues examined include the function of water molecules in providing strength and specificity of binding, the extent to which binding subsites are interdependent, and the roles of polypeptide chain flexibility and distal structural elements in contributing to specificity profiles. The studies also provide a foundation for exploring why specificity modification can be either straightforward or complex, depending on the particular system.

Keywords: enzyme kinetics; macromolecular recognition; protein engineering; protein–ligand interactions; protein structure; serine protease; site-directed mutagenesis; substrate specificity

Serine proteases were among the first enzymes to be studied extensively (Neurath, 1985). Interest in this family has been maintained in part by an increasing recognition of their involvement in a host of physiological processes. In addition to the biological role played by digestive enzymes such as trypsin, serine proteases also function broadly as regulators through the proteolytic activation of precursor proteins (Neurath, 1984; Van de Ven

et al., 1993). Examples of this regulation include the processing of trypsinogen by enteropeptidase to produce active trypsin (Huber & Bode, 1978) and the cascades of zymogen activation that control blood clotting (Davie et al., 1991). Serine proteases have also been recently shown to play essential roles in cell differentiation. For example, the *Drosophila* trypsin-like enzymes Easter and Snake are important components in the specification of ventral and lateral patterns during development (Chasan & Anderson, 1989). Asymmetry of cell fates may be the result of a protease cascade involving both of these enzymes (Smith & DeLotto, 1994).

An alternative rationale for the continued interest in serine proteases has been their emergence as one of the major paradigms for the understanding of enzymic rate enhancements and of structure–activity relationships. Until recently, all of the known enzymes fell into one of two distinct structural classes: the chymotrypsin-like and subtilisin-like families (Matthews, 1977; Fig. 1A,B). However, the crystal structure of wheat serine carboxypeptidase II (Liao & Remington, 1990; Liao et al., 1992; Fig. 1C) reveals conservation of the essential features of the catalytic apparatus within a third distinct protein fold. This homodimeric enzyme possesses the $\alpha+\beta$ fold found also in a number of other enzymes that share hydrolytic activity as their only common feature (Ollis et al., 1992). The fold consists of an 11-stranded mixed β -sheet structure surrounded by 15 helices, with the active site located at the base of a deep bowl-shaped depression in the enzyme surface (Fig. 1C).

The three serine protease classes are distinguished by the absence of any conserved secondary and tertiary motifs, but in

Reprint requests: Charles S. Craik, Departments of Pharmaceutical Chemistry and Biochemistry & Biophysics, University of California, San Francisco, California 94143-0446; e-mail: craik@cgl.ucsf.edu.

¹ Present address: Department of Chemistry and Interdepartmental Program in Biochemistry and Molecular Biology, University of California, Santa Barbara, California 93106.

Abbreviations: APP1, amyloid β -protein precursor inhibitor domain; BAP, *Bacillus alcalophilus* alkaline protease; BLAP, *Bacillus lentus* alkaline protease; BPTI, bovine pancreatic trypsin inhibitor; CMK, chloromethyl ketone; HNE, human neutrophil elastase; hGH, human growth hormone; Nva, norvaline, a linear three-carbon side chain; PAI-1, plasminogen activator inhibitor 1; pNA, *para*-nitroanilide; PPE, porcine pancreatic elastase; PROK, *Thermus albus* proteinase K; RMCP1 and RMCP2, rat mast cell proteases I and II; SBPN, *Bacillus amyloliquefaciens* subtilisin BPN'; SCARL, *Bacillus licheniformis* subtilisin Carlsberg; SGPA, *Streptomyces griseus* protease A; SGPB, *S. griseus* protease B; SGPE, *S. griseus* protease E; SSI, *Streptomyces* subtilisin inhibitor; *suc*, succinyl; *suc*-FAHY-pNA, tetrapeptide amide substrates varying at the P1 position; *suc*-XAPF-pNA, tetrapeptide amide substrates varying at the P4 position; THERM, *Thermus vulgaris* thermitase; TPA, tissue plasminogen activator. Nomenclature for the substrate amino acid residues is P_n, . . . , P₂, P₁, P₁', P₂', . . . , P_n', where P₁–P₁' denotes the hydrolyzed bond. Sn, . . . , S₂, S₁, S₁', S₂', . . . , S_n' denote the corresponding enzyme binding sites.

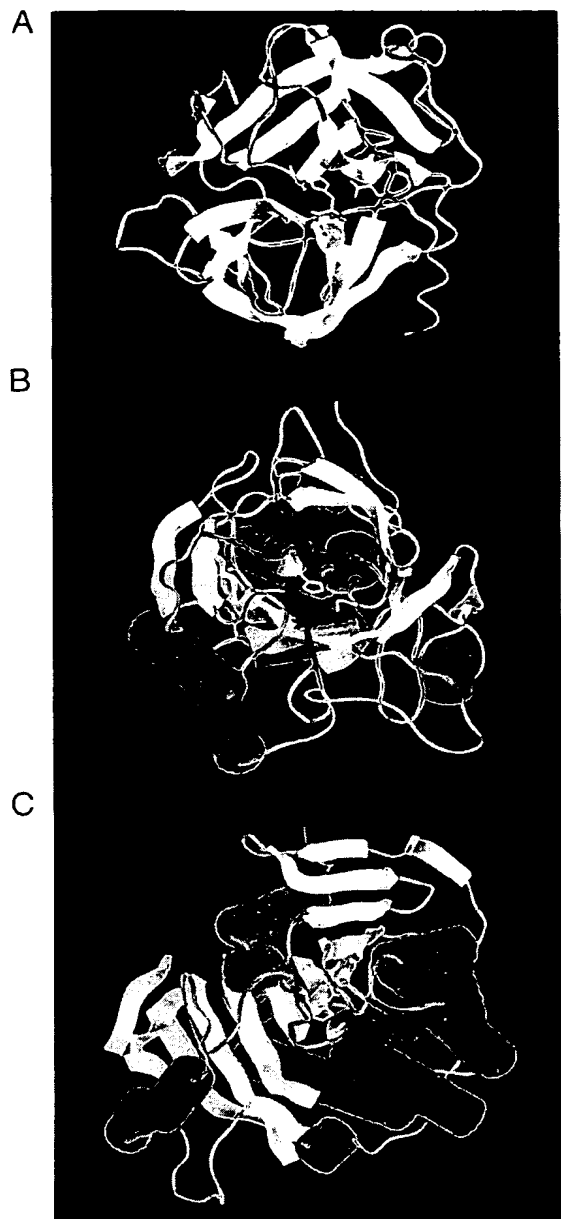


Fig. 1. Diversity of structural motifs in which the common catalytic apparatus of serine protease is embedded. Shown are ribbon drawings of chymotrypsin (A), subtilisin BPN' (B), and wheat serine carboxypeptidase (C). α -Helices are shown as red cylinders and β -strands as yellow arrows. Secondary structures were determined by the algorithm of Kabsch and Sander (1983). Each enzyme possesses two common residues of crucial importance to catalysis: a nucleophilic Ser and an adjacent His, which functions as a general base (shown in white). Enzymes are oriented identically by superposition of the backbone atoms and C β of these two amino acids. A third member of the catalytic machinery is an aspartate residue (shown at left, also in white) not conserved in position relative to the Ser and His (compare serine carboxypeptidase with the other two enzymes). Lack of conservation in position of this residue suggests that the catalytic apparatus may be better viewed as a juxtaposition of Ser-His and His-Asp dyads, rather than as a single catalytic triad.

each case, the catalytic serine and histidine residues maintain an identical geometric orientation (Fig. 1). To a lesser extent, adjacent groups that stabilize the transition state are also similarly arranged (Wright et al., 1969; Robertus et al., 1972a, 1972b; Liao et al., 1992). Thus, it appears that nature has arrived at the same biochemical mechanism by separate avenues: the chymotrypsin, subtilisin, and serine carboxypeptidase families of serine proteases are a classic example of convergent enzyme evolution (Matthews, 1977; Liao et al., 1992). The resemblance of serine carboxypeptidase to other members of the α/β -hydrolase fold family also indicates the operation of divergent evolution within this structural framework (Ollis et al., 1992). Further, a recently generated catalytic antibody has been characterized that catalyzes the stereoselective hydrolysis of norleucine and methionine phenyl esters (Guo et al., 1994). The crystal structure of this enzyme reveals the presence of a Ser-His catalytic dyad structurally similar to those of the other serine protease classes (Zhou et al., 1994). A similar catalytic mechanism is therefore suggested, indicating that the antibody fold may well be a fourth structural framework capable of supporting proteolytic activity in a serine protease-like fashion.

We consider here the structural and kinetic basis for the diversity of substrate specificity in the subtilisin and chymotrypsin-class serine proteases. Emphasis is placed on those systems for which both crystallographic and detailed kinetic measurements are available. After a brief review of the common mechanism of the three classes and the role of mutational analysis in its further elucidation, we concentrate much of our attention on the three enzymes subtilisin BPN', α -lytic protease, and trypsin. In each case, an extensive structure-function analysis has been applied to address the roles of particular amino acids in contributing to the observed specificity profiles. The wealth of information available on the chemical and kinetic mechanisms of catalysis and the large data base of homologous sequences provide an essential framework that supports these studies. Although the functional and/or structural properties of many of the mutant proteases can be given a relatively straightforward and objective description, there are also many examples where the data cannot be easily encapsulated. In these cases, some subjectivity in the description of kinetic and structural parameters is unavoidable, and other interpretations of the same data could yield different overall conclusions.

The catalytic mechanism

The vast majority of early studies on the serine proteases focused on the elucidation of the chemical and kinetic mechanisms of catalysis (reviewed by Bender & Killheffer, 1973; Blow, 1976; Kraut, 1977; Polgar, 1989). Hydrolysis of ester and amide bonds proceeds by an identical acyl transfer mechanism in all enzymes of the subtilisin and trypsin families (Fig. 2A,B,C). Michaelis complex formation is followed by attack on the carbonyl carbon atom of the scissile bond by the eponymous serine of the catalytic triad, which is enhanced in nucleophilicity by the presence of an adjacent histidine functioning as a general base catalyst. Proton donation by the histidine to the newly formed alcohol or amine group then results in dissociation of the first product and concomitant formation of a covalent acyl-enzyme complex. The deacylation reaction occurs via the same mechanistic steps, with the attacking nucleophile provided by a water molecule that approaches from the just-vacated leaving group

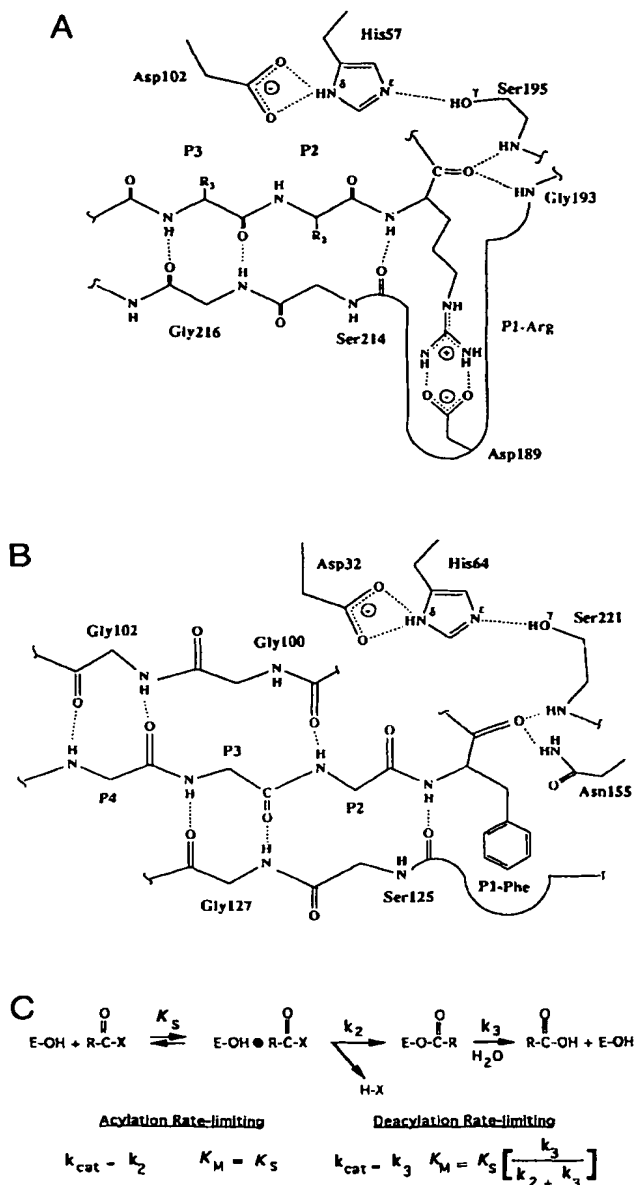


Fig. 2. Chemical and kinetic mechanisms of catalysis for serine proteases. The catalytic groups of trypsin (**A**) and subtilisin (**B**) are shown interacting with an oligopeptide substrate binding to the P1–P4 sites. (Nomenclature for the substrate amino acid residues is $P_n, \dots, P_2, P_1, P_1', P_2', \dots, P_n'$, where P1–P1' denotes the hydrolyzed bond. $S_n, \dots, S_2, S_1, S_1', S_2', \dots, S_n'$ denote the corresponding enzyme binding sites [Schechter & Berger, 1968].) Note the distinction in residues that form the oxyanion hole; in subtilisin, part of the interaction is made by an enzyme side chain. The binding site for the oligopeptide also differs; in subtilisin it forms the central strand of a three-stranded antiparallel β -sheet. The S1 site of trypsin and the S1 and S4 sites of subtilisin are the major sites where mutagenesis has been used to probe specificity. **C:** Common kinetic mechanism of catalysis for serine proteases indicating the meaning of the mechanistic rate constants and their relationship to the Michaelis parameters. The correct interpretation of k_{cat} and K_M differs depending on the rate-limiting step in catalysis, which varies among the different enzymes as well as among differing substrates of the same enzyme.

side. Each step proceeds through a tetrahedral intermediate, which resembles in structure the high-energy transition state for both reactions. This mechanism is capable of accelerating the rate of peptide bond hydrolysis by a factor of more than 10^9 relative to the uncatalyzed reaction (Kahne & Still, 1988).

Extensive structural evidence obtained from X-ray crystallographic and NMR investigations has provided conclusive corroboration of the essential features of this mechanism (reviewed by Steitz & Shulman, 1982). The investigations have been favored by the availability of good ground-state and transition-state substrate analogs, which have been used to obtain high-resolution images of these interactions. The scissile bond of the substrate is bound directly adjacent to the Ser–His catalytic couple in all the complexes studied. A strong hydrogen bond between these two amino acids, necessary to subsequent proton transfer, is formed only after substrate is bound. A binding site for the oxyanion of the intermediate is formed by the Gly 193 and Ser 195 backbone amide nitrogens in the chymotrypsin-like enzymes (Fig. 2A), by one amide nitrogen and the Asn 155 side chain in the subtilisin family (Fig. 2B), and by the backbone amides of Tyr 147 and Gly 53 in the serine carboxypeptidases (Liao et al., 1992). The interactions made in the S1–S4 enzyme sites (see Fig. 2 legend for substrate nomenclature) by the P1–P4 positions of substrate form an antiparallel β -sheet hydrogen bonding arrangement in the chymotrypsin and subtilisin families. Because the active site of wheat serine carboxypeptidase II does not possess similarly exposed peptide backbone groups, it seems likely that substrate binding N-terminal to the scissile bond will occur in a different fashion in this family (Liao et al., 1992). Another unique structural feature of carboxypeptidase is an extensive hydrogen bonding network, which interacts with the C-terminal carboxylate of the substrate, essential to its activity as an exopeptidase (Mortenson et al., 1994).

Mutational analysis of both subtilisin and trypsin has confirmed the essential roles of Ser 195 and His 57 in providing rate acceleration. Replacement of the catalytic Ser 221 and His 64 residues of subtilisin with alanine results in decreases of 10^4 – 10^6 -fold in k_{cat} (Carter & Wells, 1987, 1988). A decrease of 10^6 -fold when the two residues are simultaneously replaced with alanine showed that the two catalytic moieties function in a highly cooperative manner: mutation of either component reduces activity to a baseline level. Similar results were obtained by analogous mutations of Ser 195 and His 57 in rat trypsin (Corey & Craik, 1992). This study also showed that enzyme variants such as H57K and H57E, which might provide an alternative general base, were ineffective, further underscoring the importance of the native catalytic triad geometry. These experiments, as well as others involving replacement of Ser 195 with a Cys (Higaki et al., 1989; McGrath et al., 1989) and engineering a metal-actuated activity switch involving His 57 (Higaki et al., 1990; McGrath et al., 1993), clarify the role of these active-site moieties. The mutational data are in agreement with early chemical modification experiments, which also indicated that Ser 195 and His 57 play crucial roles in catalysis (Dixon et al., 1956; Shaw et al., 1965).

The residual activity remaining in subtilisin after removal of the catalytic moieties was attributed to remaining binding determinants that stabilized the transition state complex. One such determinant is provided by a hydrogen bonding interaction of Asn 155 with the oxyanion intermediate. Mutation of Asn 155 to a variety of other amino acids resulted in 10^2 – 10^3 -fold de-

creases in k_{cat}/K_m (Bryan et al., 1986; Wells et al., 1986; Carter & Wells, 1990). This provides support for the proposals made on the basis of crystallographic studies, which suggested that a weak hydrogen bond to Asn 155 in the Michaelis complex is strengthened in the transition state (Robertus et al., 1972b; Poulos et al., 1976). Interestingly, mutation of Thr 220 of subtilisin showed that it stabilizes the transition state by 2 kcal/mol despite the fact that the side-chain O^γ lies 4.0 Å from the oxyanion, too far for a direct interaction (Braxton & Wells, 1991). One explanation for the influence of Thr 220 was proposed to be that dynamic fluctuations of the protein structure (Rao et al., 1987) cause transient direct interactions to occur. An alternative suggestion was that the oriented Thr 220 side-chain dipole may stabilize the transition state at a distance, by influencing the electrostatic potential in the active site. Significant perturbation of the pK_a of the catalytic His 64 results from mutation of charged surface residues some 12–20 Å distant from the active site (Russell et al., 1987; Loewenthal et al., 1993). Similar mutation of distant charged residues affects the stability of complex formation with a transition-state analog inhibitor (Jackson & Fersht, 1993). These observations support the hypothesis that long-range electrostatic interactions may play a small but significant role in stabilizing the catalytic transition state.

Considerable controversy has surrounded the role of an additional component of the catalytic apparatus, a conserved buried aspartate residue first described in the crystal structure of chymotrypsin (Matthews et al., 1967; Blow et al., 1969). Mutation of this residue confirmed its essential role, because all variants of trypsin and subtilisin in which the aspartate is absent are decreased in catalytic efficiency by at least a factor of 10^4 (Craik et al., 1987; Sprang et al., 1987; Carter & Wells, 1988; Corey & Craik, 1992). The early suggestion of a two-proton transfer model, in which the Asp accepts a proton to become uncharged in the transition state, now appears to be unsupported by the bulk of the experimental (Bachovchin & Roberts, 1978; Markley, 1979; Kossiakoff & Spencer, 1981) as well as theoretical (Warshel et al., 1989) evidence. One role for the conserved Asp appears to be ground-state stabilization of the required tautomer and rotamer of the catalytic His (Craik et al., 1987; Sprang et al., 1987). In addition, because the His imidazole ring acquires a proton in the transition state, the Asp carboxylate can provide compensation for the developing positive charge. Its role may therefore be considered similar to that of the hydrogen bond donor groups in the oxyanion hole, which compensate the developing negative charge on the substrate carboxyl oxygen atom (Warshel et al., 1989; Fig. 2A,B). Experimental evidence for the role of electrostatic stabilization of the trypsin transition state has been obtained by mutation of the conserved Ser 214, which forms a solvent-inaccessible hydrogen bond to Asp 102, to various charged and uncharged amino acids (McGrath et al., 1992). Decreases in the free energies of catalysis were in agreement with electrostatic calculations, based on crystal structures of the mutants, which predicted these losses of activity.

Comparative analysis of the structures of chymotrypsin, subtilisin, and serine carboxypeptidase shows that the precise geometric orientation of the Asp is not conserved relative to the Ser–His catalytic diad (Liao et al., 1992; compare Fig. 1A,B,C). In contrast to chymotrypsin and subtilisin, the plane of the Asp carboxylate in carboxypeptidase is tilted far out of the plane of the His imidazole, such that the His–Asp hydrogen bond is 45°

out of the carboxylate plane. This geometry is unfavorable for proton transfer from His to Asp and provides further evidence against the double proton-transfer mechanism. A detailed analysis of high-resolution subtilisin structures also showed differences in the Asp–His hydrogen bonding relative to trypsin (McPhalen & James, 1988). It now appears that the Asp can occupy virtually any position relative to the Ser–His diad. Therefore, it may be more accurate to regard the operation of the serine protease catalytic machinery as two diads—Ser–His and His–Asp—that operate in concert, rather than as a single catalytic triad (Liao et al., 1992). In this context, it is of interest to note that relocation of the Asp 102 carboxylate group to position 214 in trypsin significantly reconstitutes the activity lost in the variants D102S and D102N (Corey et al., 1992). The crystal structure of this mutant shows that Asp 214 still interacts with His 57, but in an altered geometric orientation in which the plane of the carboxylate is displaced from that of the imidazole ring by 40°. The relatively high catalytic efficiency of this variant thus supports the view of the catalytic apparatus as a juxtaposition of two diads.

Substrate specificity in the subtilisin family

The catalytic machinery and substrate binding clefts of the subtilisin-class serine proteases are embedded in a single-domain molecule (Wright et al., 1969; McPhalen & James, 1988). Six crystal structures are available in this family: *Bacillus amyloliquefaciens* subtilisin BPN' (Novo) (Wright et al., 1969; McPhalen & James, 1988), *Bacillus licheniformis* subtilisin Carlsberg (Bode et al., 1986a; McPhalen & James, 1988), *Thermus vulgaris* thermitase (Gros et al., 1989), *Thermus albus* proteinase K (Betz et al., 1988), *Bacillus lentus* alkaline protease (Betz et al., 1992), and *Bacillus alcalophilus* alkaline protease (van der Laan et al., 1992). The central core of the globular heart-shaped molecule is formed by a seven-stranded parallel β -sheet (Fig. 1B). Nine α -helices are packed against the sheet in a mostly antiparallel fashion relative to the β -strands; seven of these are on the same face and form the larger of two subdomains defined on either side (McPhalen & James, 1988). A two-stranded antiparallel β -sheet is also formed in the larger subdomain near the C-terminus of the chain. The active site is located in the larger subdomain adjacent to the central β -sheet; the catalytic Ser 221 is found near the amino-terminus of a long α -helix, which follows the small antiparallel sheet (Fig. 1B; McPhalen & James, 1988; numbering system for SBPN is used throughout).

Nearly all of the secondary structure elements of the enzymes are very highly conserved. A central core of 194 amino acids has been defined by comparison of the known structures, which contains nearly all of the conserved α -helices and β -strands (Siezen et al., 1991). The fungal-derived PROK deviates most significantly in structure but still superimposes these equivalent C α atoms with RMS deviation of about 0.9 Å (the other prokaryotic enzymes superimpose at 0.4 Å to 0.65 Å; Siezen et al., 1991). If PROK is omitted, a more extended core of 232 amino acids can be defined among the bacterial species of known structure. An extensive sequence comparison of 47 subtilisin-class enzymes showed a subdivision into two subclasses, based on conserved differences in certain parts of the alignment. SBPN, SCARL, THERM, BAP, and BLAP are members of subclass I; the structurally divergent PROK is a representative of subclass II (Siezen et al., 1991). Although the homologous catalytic core of some

270 amino acids is found in all subtilisins, some of the enzymes possess large insertions in this domain, and many also possess C-terminal extensions resulting in polypeptide chains as long as 1,775 amino acids. This large database of sequence information forms the basis for homology modeling of those enzymes for which no tertiary structure is available (Siezen et al., 1991, 1993).

Crystal structures of enzyme-inhibitor complexes have identified substrate binding determinants extending over nine amino acids, from P6 to P3'. The structures include several peptide chloromethyl ketone complexes, in which subsites P1-P3 are occupied (Robertus et al., 1972a; Poulos et al., 1976), as well as complexes of SCARL with the protein inhibitor eglin C (Bode et al., 1986a; McPhalen & James, 1988), SBPN with eglin C, chymotrypsin inhibitor 2 and *Streptomyces* subtilisin inhibitor, (Bode et al., 1986a; McPhalen & James, 1988; Takeuchi et al., 1991a, 1991b), THERM complexed to eglin C (Gros et al., 1989), and PROK complexed with peptide inhibitors (Betz et al., 1993). In each of these complexes, the inhibitor chain binds in a surface channel of the enzyme, which accommodates six residues from P4 to P2'. On the N-terminal side of the scissile bond, the P1-P4 residues of the substrate main chain are invariably inserted between two β -strands of the enzyme at positions 125-127 and 100-102 (Fig. 2B). The substrate thus forms the central strand of a three-stranded antiparallel sheet unique to the subtilisins; in the chymotrypsin-like proteases, this structure is not formed because only the strand corresponding to residues 125-127 is present (Fig. 2A).

Subtilisins in general show broad substrate specificity profiles and often display a preference for large hydrophobic groups at position P1 (Markland & Smith, 1971). At this position specificity arises from a broad open S1 binding cleft formed on one side by the two β -strands, which interact with the P1-P4 substrate residues, and on the other by a loop comprising residues 155-166 (Fig. 3). This loop varies in size among members of the family (Siezen et al., 1991). In SBPN, two different modes of binding exist to accommodate either P1-Phe or P1-Lys substrates (Robertus et al., 1972a; Poulos et al., 1976). The Phe ring binds deeply in the S1 cleft near Gly 166, whereas the charged Lys extends across the cleft to form a salt bridge with Glu 156. A prominent hydrophobic cavity is also present for binding of the P4 substrate side chain (Fig. 3). These two sites have been the focus of much of the work on substrate specificity. Interactions made in the more distal sites influence catalytic efficiency markedly, and there is evidence for nonadditivity of mutational effects suggesting a functional communication between sites (Grøn & Breddam, 1992).

Interactions in the S1 site

The most intensively studied member of the subtilisin family is SBPN, which has been the subject of extensive protein engineering investigations (reviewed in Wells et al., 1987b; Wells & Estell, 1988). The enzyme efficiently cleaves peptidyl amide substrates possessing a broad range of P1 amino acids, with the k_{cat}/K_m value showing a linear dependence on the hydrophobicity of the substrate side chain. The preference of the enzyme at this position is roughly Tyr, Phe > Leu, Met, Lys > His, Ala, Gln, Ser > Glu, Gly (Estell et al., 1986; Wells et al., 1987c). To investigate the role of hydrophobicity more closely, 12 different amino acids were substituted for Gly 166, which lies at the base of the pocket (Fig. 3). Analysis of the mutants showed that an increase in the

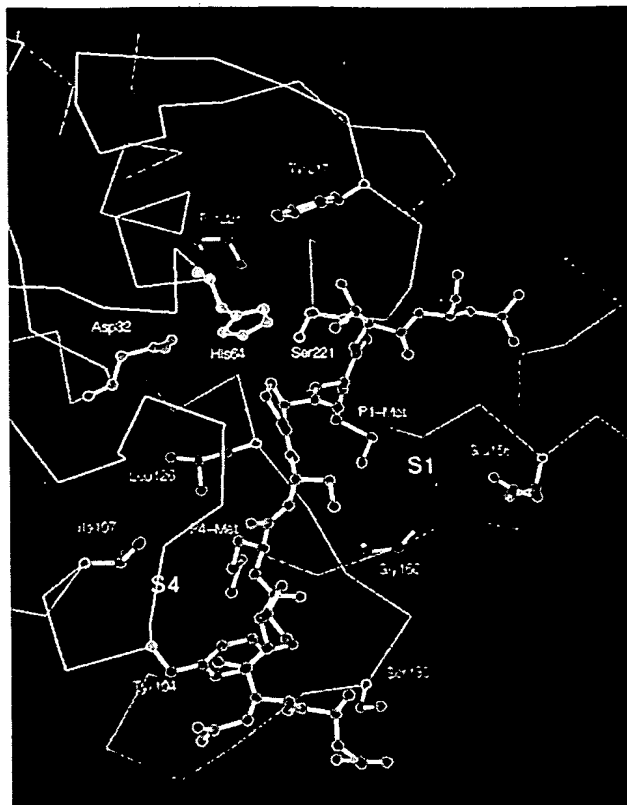


Fig. 3. Structure of the S1 and S4 sites of subtilisin BPN' showing binding of a peptide derived from the cocrystal structure with *Streptomyces* subtilisin inhibitor. An α -carbon trace of the protein is shown in thin blue lines. Catalytic residues are in yellow, and the inhibitor chain is in green with the P1 and P4 side chains labeled in blue. Locations of amino acids at which the S1 and S4 sites have been mutated are indicated in red. In the subtilisin family, both the S1 and S4 sites are generally specific for hydrophobic side chains, but Glu 156 in the S1 site of subtilisin BPN' provides activity toward P1-Lys side chains as well. At both sites, specificity alteration is readily achievable by the substitution of a small number of residues directly in contact with substrate. Modulation of the hydrophobic specificity profiles has been achieved at both sites, and altered specificity toward charged residues has been achieved in the S1 pocket.

side-chain volume at this position, which consequently decreases the size of the S1 cleft, caused substantial reductions (up to 5,000-fold) in k_{cat}/K_m toward large P1 amino acids. This presumably occurs due to steric repulsion, which predominates over the favorable effect of a more hydrophobic pocket. Catalytic efficiencies toward small P1 side chains were increased by up to 10-fold in these variants. An optimal combined volume for the S1 and P1 side chains of 160 Å³ was estimated from these data. It appears that hydrophobicity of the S1 site is the main driving force for specificity, whereas other effects, such as attractive van der Waals forces and hydration of polar side chains, have a lesser though still significant role.

Because these studies showed that specificity is easily modulated by replacing amino acids directly contacting substrate, it seemed plausible that more distant portions of the enzyme struc-

ture might be of little importance. This idea was further explored by a mutational study in which several amino acids from the related SCARL enzyme were exchanged for those in SBPN (Wells et al., 1987a). Although these two enzymes differ by 31% in sequence, only three substitutions lie within 7 Å of the S1 pocket. Two of these, at positions 156 and 217 (Fig. 3), directly contact substrate (residue 217 is in the S1' site). A third residue at position 169 is positioned behind the loop comprising residues 156–166, which forms one side of the S1 pocket. In SBPN the amino acids are Ser 156, Ala 169, and Leu 217; these replaced the analogous Glu 156, Gly 169, and Tyr 217 of SCARL. The wild-type enzymes differ by factors of 6–60-fold in their k_{cat}/K_m values toward peptidyl amide substrates possessing P1-Glu, Met, Phe, Gln, or Ala; in each case, SBPN is more efficient (Wells et al., 1987a).

The triple mutant E156S/G169A/Y217L was found to exhibit a substrate specificity profile very similar to that of SCARL. Cleavage at each of the P1 amino acids tested occurred with efficiencies within threefold of the target protease (Wells et al., 1987a). These data demonstrate that, of the 86 amino acid differences between the two enzymes, three alone are largely sufficient to determine the differences in specificity. Further, analysis of singly and doubly substituted variants showed that the E156S mutation is alone almost entirely responsible for the shift in specificity profile. Because the activity of the E156S/Y217L enzyme was found to be within twofold of the triple mutant, it appears P1 substrate specificity is in fact locally determined to a significant degree.

The behavior of the E156S variant is similar to that of other mutant SBPN enzymes also possessing electrostatic substitutions in the S1 site (Table 1; Wells et al., 1987c). Sixteen variants were constructed at positions 156 and 166, each of which altered the electrostatic potential of the S1 site by introducing or removing Arg, Lys, Glu, or Asp residues at one or both sites. Analysis of the mutants showed that increases as high as 10^3 -fold in k_{cat}/K_m toward complementary charged substrates could be achieved. To assess the contribution of electrostatic free energy to the stabilization of the transition-state complex, parallel substitutions of roughly isosteric but uncharged residues (Met replacing Lys; Gln replacing Glu) were also made. For example, it was found that increasing the positive charge in the S1 site increases k_{cat}/K_m much more for P1-Glu than for P1-Gln sub-

strates. In this way, substrate binding effects associated solely with the charge-charge interaction could be isolated.

Several of the S1-site specificity variants were also utilized in a different study that addressed the ability of SBPN to function as a peptide ligase (Abrahmsen et al., 1991). This reaction occurs when peptides bearing a free amino-terminal group can compete effectively with water for attack on the acyl-enzyme intermediate. The intrinsic low level of ligase activity normally present in SBPN was enhanced by substitution of the active-site Ser 221 by Cys, which shifts the relative preference toward aminolysis by more than 10^3 -fold (Nakatsuka et al., 1987). The additional mutation P225A improves ligase activity by an additional 10-fold (Abrahmsen et al., 1991). The usefulness of this SBPN variant (referred to as subtiligase) for the synthesis of proteins was improved by introducing specificity variants G166I, G166E, and E156Q/G166K into the S221C/P225A framework. Preferred ligation of P1-Glu, P1-Phe, P1-Lys, and P1-Arg esters was achieved; the specificity for ligation mirrored that for cleavage of peptidyl amide substrates (Estell et al., 1986; Wells et al., 1987c). The ability to modulate the S1-site specificity thus provides greater flexibility in the choice of ligation junctions. Subtiligase has been used to synthesize ribonuclease A and active-site variants of this enzyme by stepwise ligation of six esterified peptide fragments 12–30 residues long (Jackson et al., 1994).

Substrate-assisted catalysis

Substrate-assisted catalysis represents a strategy for enhancing the specificity of proteolytic cleavage. Subtilisins lacking the catalytic His 64 can be reconstituted by including a histidine residue within the substrate (Carter & Wells, 1987; Carter et al., 1989, 1991). By placing a His at the P2 position of peptidyl amide substrates, specificity of up to 400-fold was achieved relative to analogous P2-Gln and P2-Ala substrates. The increased specificity at position P2 occurs within the context of a compromised enzyme: H64A subtilisin is reduced 10^6 -fold in k_{cat}/K_m , and H64A in the presence of a P2-His substrate remains 5,000-fold less efficient than the wild-type enzyme (Carter & Wells, 1987). Mutation of Ser 221, Asp 32, and Asn 155 in the context of H64A suggested that interactions of the catalytic His with the Ser and Asp residues are severely compromised when the His is present in the substrate (Carter et al., 1991). By contrast, the oxyanion hole interactions appear much less disrupted. Model-building of P2-His substrates indicates that the imidazole ring can occupy roughly the same position as that of His 64 in the native enzyme, although some deviation in hydrogen bond distances and angles exists, which may partially explain the reduced activity.

The large database of S1-site specificity variants was again used to enhance the selectivity of proteolytic cleavage by the prototype H64A enzyme (Carter et al., 1989). For example, an improvement of 20-fold in cleavage of *suc*-FAHY-*pNA* was observed by introducing the S1 and S1'-site mutations E156S, G169A, and Y217L (Estell et al., 1986; Wells et al., 1987c), which increase catalytic efficiency toward P1-Phe and P1-Tyr substrates. The additional mutation G166A enhanced specificity for P1-Phe but not P1-Tyr substrates, as expected because the C^β of Ala 166 appears to cause steric hindrance to the binding of the larger Tyr side chain. Little specificity was observed on the

Table 1. Engineering electrostatic interactions in subtilisin^a

	Net charge	P1-Glu	P1-Lys
E156D166	-2	—	16,200
E156N166	-1	40	17,800
E156Q166	-1	16	12,600
S156D166	-1	17	17,400
E156G166 (wt)	-1	35	39,800
Q156G166	0	620	1,070
Q156N166	0	110	5,600
E156R166	0	810	1,550
Q156K166	+1	66,000	1,700
S156K166	+1	16,200	5,400

^a Substrate: *suc*-Ala-Ala-Pro-Glu/Lys-*pNA*. k_{cat}/K_m , s⁻¹ M⁻¹.

C-terminal side of the peptide bond in the cleavage of peptide substrates. The mutant subtilisins have been shown to selectively cleave designed target sites in fusion proteins, even under adverse conditions, making them a useful additional tool in the repertoire of protein chemists (Carter et al., 1989).

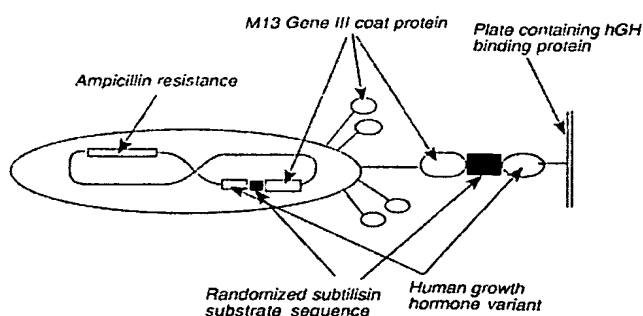
Further insight into substrate-assisted catalysis was provided by a novel approach using phage display technology (Matthews & Wells, 1993; Fig. 4A). A randomized target substrate sequence for an improved H64A subtilisin (Carter et al., 1989) was inserted between an amino-terminal affinity domain representing a variant of human growth hormone, and the carboxy-terminal domain of the M13 phage gene III coat protein. A collection of phage particles bearing different substrate sequences is bound to immobilized hGH-binding protein and cleaved by subtilisin, so that phage bearing good substrate sequences are eluted and those bearing poor sequences remain bound. Propagation of the phage further enriches for efficient or inefficient cleavage sites. Analysis of the sequences that were efficiently cleaved revealed that P1'-His as well as P2'-His-containing substrates could function in substrate-assisted catalysis. Analysis of cleavage of fusion proteins linked to alkaline phosphatase, which provides an easily assayed activity, suggested that P1'-His-mediated cleavage was comparable in efficiency to P2'-His cleavage. Further study of P1'-His cleavage would be informative because release of the leaving group after formation of the acyl-enzyme implies that no catalytic His is present to assist in deacylation. Molecular modeling has shown that a P1'-His can also occupy the position vacated by His 64 in an H64A variant (Matthews & Wells, 1993).

The P4-S4 interactions

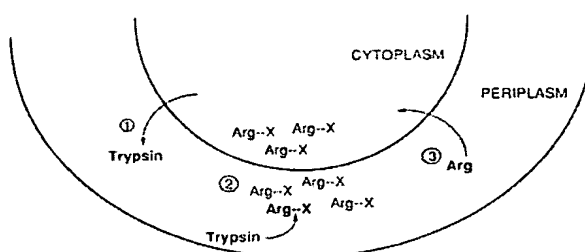
Considerable specificity toward substrate residues distant from the scissile bond exists in the subtilisin-class family. A thorough mapping of the preferences of two enzymes—SBPN and BLAP—shows that the most marked distal interaction occurs on the N-terminal side of the substrate at the S4 enzyme site (Grøn et al., 1992). Mutational analysis at this position has been applied to three of the enzymes of known structure: SBPN (Eder et al., 1993; Rheinacker et al., 1993, 1994), BLAP (Bech et al., 1992, 1993; Sørensen et al., 1993), and BAP (Teplyakov et al., 1992). The S4 site is formed from the juxtaposition of two structural elements: residues 100–107 at the amino-terminus of an α -helix in the small subdomain and residues 125–132 in an adjacent surface loop. Substrate interactions include both the main-chain β -sheet hydrogen bonds as well as contacts with the side chains of residues 104, 107, 126, and 135, which line the sides and base of the site (Fig. 3). Of the amino acids shaping the cleft, only Gly 127 is invariant in the family (Siezen et al., 1991).

In SBPN, the amino acid side chains in the S4 site are Tyr 104, Ile 107, and Leu 126, which create a large hydrophobic pocket. Accordingly, the substrate preferences follow the series Phe > Leu, Ile, Val > Ala for cleavage of peptidyl amide substrates (Rheinacker et al., 1993). Slightly different preferences following the same general trend were observed toward long peptides occupying subsites S5–S5' (Grøn et al., 1992). However, the range of k_{cat}/K_m values varies only over a three- to sixfold range. It was suggested that the small variability might be due to compensatory shrinkage of the S4 site upon binding of smaller side chains (Takeuchi et al., 1991a). Efficiencies toward polar resi-

A Protease substrate phage selection



B Selection for active trypsin mutants



C Phage display of trypsin

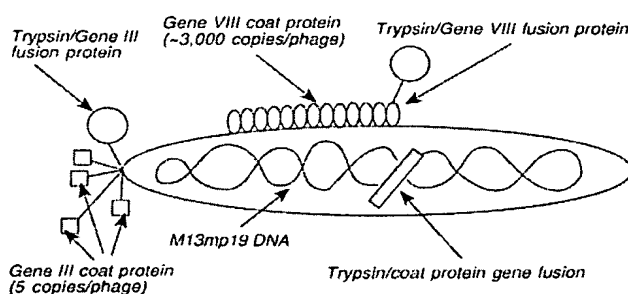


Fig. 4. Randomization methodologies employed in isolation of serine protease substrate specificity mutants. **A:** "Substrate phage" approach applied to subtilisin. In this method, the sequence of the substrate rather than the enzyme is varied to explore the substrate specificity at many of the subsites. By using H64A subtilisin as the cleaving protease, it was discovered that substrate-assisted catalysis functions when the substrate His is present at the P1' as well as the P2 position. Note that in phage display systems, the phage particle provides a "package" in which the mutant DNA and variant protein are physically linked. This facilitates analysis after enrichment of those phage bearing good substrate sequences. **B:** Genetic selection for the isolation of trypsin variants. Periplasmic expression of a variant trypsin capable of cleaving the nonnutritive Arg-X substrate (1, 2) leads to release of free Arg (3), which enters the cytoplasm and relieves auxotrophy. Twenty variant tryptins possessing altered Arg/Lys specificity ratios have been isolated in this manner. **C:** Phage display approach for the isolation of trypsin variants. A wild-type trypsin gene fused to the M13 gene III coat protein specifically binds immobilized ecotin, a dimeric protein inhibitor of mammalian serine proteases that is found in the bacterial periplasm.

dues are decreased by more than 100-fold relative to hydrophobic amino acids (Grøn et al., 1992).

Tyr 104, Ile 107, and Leu 126 were mutated singly and in combination to amino acids that in every case were smaller than the wild-type residue. The following variant enzymes were characterized kinetically toward amide substrates of the form *suc*-XAPF-*p*NA: Y104F, Y104A; I107G, I107A, I107V; L126G, L126A, L126V, and the double mutants I107G/Y104A, I107G/L126A, I107G/L126V (Rheinnecker et al., 1993, 1994). These alterations test the effects of enlarging the P4 pocket as well as the consequences of deleting a hydrogen bond present between the side chains of Tyr 104 and Ser 130.

It was found that the Tyr 104-Ser 130 hydrogen bond has little effect on enzyme efficiency or specificity: Y104F SBPN hydrolyzes P4-Ala, Val, Ile, Leu, and Phe substrates nearly identically to the wild-type enzyme. The effect of introducing Ala at this position is similar to that caused by decreasing the size of Ile 107: in each case specificity is increased for residues possessing large side chains at P4. Among the single mutants at positions 104 and 107, the largest improvements in the relative specificity for P4-Phe relative to P4-Ala are roughly 200-fold for both Y104A and I107G. For these variants, the effects are achieved by maintaining approximately wild-type levels of k_{cat}/K_m toward Phe and sharply decreasing efficiencies toward Ala and the other smaller substrate residues. Mutation of Leu 126 had smaller effects on relative specificities, but large decreases in the range of 10–10⁴-fold were observed in k_{cat}/K_m , with decreased efficiency correlated with decreasing size of the side chain.

The three double mutants also showed strong preference for large side chains at position P4 (Rheinnecker et al., 1994). Among these enzymes, the mutant I107G/L126V improves the P4-specificity for large side chains to 340-fold relative to P4-Ala, but in this case the maximal discrimination was achieved with P4-Leu rather than P4-Phe. The other two double mutants similarly exhibited a maximal preference for P4-Leu. In all cases, nonadditivity was observed relative to the single mutants, as expected from the close proximity of the three side chains. Kinetic parameters were also measured toward the single-residue substrate acetyl-tyrosine ethyl ester, which might be considered as a probe measuring the extent to which S4-site mutants affect the functioning of the S1 site. Large decreases of up to 60-fold were observed, with the largest effects occurring for the double mutants. However, the same variants exhibit comparable efficiencies to wild-type when measured toward favored *suc*-XAPF-*p*NA substrates. This suggests that less productive binding may occur in the absence of the subsite interactions, particularly because the ester substrate is more easily cleaved owing to the better leaving group.

The substrate preference of BLAP at the P4 substrate position is also toward large hydrophobic side chains (Grøn et al., 1992). A broader range of specificities exists than in SBPN: in this case, a 24-fold (rather than sixfold) increase in k_{cat}/K_m when progressing from small to large hydrophobic amino acids is observed. The individual subsite interactions do not affect the overall catalytic efficiencies in an additive manner, suggesting that functional communication occurs and is mediated by structural elements of the protein (Grøn & Breddam, 1992). For example, modest substrate preferences at some sites are masked if the optimal P1-Phe and/or P4-Phe residues are present. These amino acids dominate the cleavage efficiency such that an up-

per limit in k_{cat}/K_m is reached even when other subsites are filled by nonpreferred residues. These other sites are therefore less important when a good substrate rather than a poor substrate is bound. This study underlines an important principle: optimal subsite mapping of subtilisins (and other proteases) should be carried out using sets of matched substrates where the interdependency of binding sites is not manifested. In the case of BLAP, the presence of an anthraniloyl group at P5 and a Pro at P2 apparently disrupts the P1-Phe and P4-Phe interactions, such that a substrate series containing these nonoptimal groups permits distribution of P1' site preferences over a 15-fold range. Only a 50% difference between the most and least favored P1' amino acid is observed in the absence of the nonoptimal groups, which prevents accurate mapping of the true subsite preference (Grøn & Breddam, 1992).

The structure of the BLAP S4 pocket is similar to that of SBPN. The side chains of Val 104, Ile 107, Leu 126, and Leu 135 form the base and one side of the pocket, whereas Ser 128, Ser 130, and Ser 132 are situated along the outside rim with each of the side-chain hydroxyl groups pointing inward. The substitution of Val 104 for the Tyr present in SBPN allows Leu 135 access to the substrate in BLAP. The only other difference in the pocket between the two enzymes is the presence of Gly 128 rather than Ser 128 in SBPN. A total of 21 mutants in the BLAP S4 site have been constructed and analyzed (Bech et al., 1992, 1993; Sørensen et al., 1993). At position 104 it was found that bulky hydrophobic side chains produced enzymes that preferentially cleaved small hydrophobic side chains, and conversely, smaller amino acids increased specificity toward large substrates. This behavior is reminiscent of the effects caused by increasing the size of residue Gly 166 in the S1 site of SBPN (Estell et al., 1986; see above). Mutations at other positions in the BLAP S4 site often also showed these effects, but in many cases complex specificity profiles not immediately interpretable in simple terms were obtained. What does appear clear is that both steric and hydrophobic effects play important roles in determining the S4 specificity profile (Bech et al., 1993; Sørensen et al., 1993). For some mutants it was further suggested that structural flexibility is also critical.

Distinguishing the degree to which hydrophobicity, steric exclusion, and substrate-induced conformational changes function to determine specificity profiles requires high-resolution structural information on the mutant enzymes. Such information has begun to be obtained in the study of BAP variants (Teplyakov et al., 1992). Substitution of Val 104 in this enzyme with Trp increased activity toward *suc*-AAPF-*p*NA by 12-fold. The crystal structure of the uncomplexed variant showed that no other structural change occurs and that the S4 site is now blocked off such that a modeled P4-Ala residue makes a good van der Waals contact with Trp 104. Trp 104 in this variant is oriented nearly identically to Trp 104 in THERM, which also exhibits high activity toward *suc*-AAPF-*p*NA.

Comparison of the structures of SSI and a P4-Met to Gly mutant of SSI complexed to SBPN showed that the S4 site undergoes a substantial shrinkage upon binding of P4-Gly (Takeuchi et al., 1991b). The structural flexibility in this enzyme raises the possibility that a capacity for such rearrangement may exist in other members of the family as well. Required for an assessment of the degree of flexibility, and the extent to which amino acid alterations affect this property, are crystal structures of wild-type and mutant enzymes complexed to substrate analogs pos-

sessing small and large side chains at the P4 position. In the case of BAP, for example, it would be of interest to determine the catalytic efficiencies of the wild-type and V104W enzymes toward larger hydrophobic P4-side chains and then to carry out a systematic structural analysis of complexes of each enzyme with analogous inhibitors. Such an analysis for the chymotrypsin-like α -lytic protease has yielded substantial insight into the structural basis for enzyme flexibility (Bone et al., 1991; see below).

Together these mutational alterations within the subtilisin S1 and S4 sites allow two important conclusions: (1) only the local environment of amino acids directly contacting substrate need be considered in designing specificity changes; (2) there is no important distinction between hydrophobic and polar enzyme-substrate interactions because each type is manipulatable to generate new specificity profiles while maintaining high activity. The importance of these generalizations to protein design in other systems depends upon the extent to which the structural design of the binding cleft, and the nature of the reaction being catalyzed, are crucial parameters. As we shall see, structural context can have great influence in mediating the extent to which specificity alteration is straightforward. A clue to its important role can be seen in the dependence of catalytic efficiency on the extent to which subsites are filled. The signal that distal portions of substrate are bound is transmitted over large distances and must in some way be mediated by the intervening protein structure. Long-range effects are key in the chymotrypsin family of enzymes, both in terms of filling subsites as well as in determining specificity at a single site (Corey et al., 1992; Hedstrom et al., 1992, 1994a, 1994b; Perona et al., 1995; see below).

Prohormone convertases: Specificity toward paired dibasic residues

Tissue-specific processing of precursor proteins in mammalian cells is accomplished by a subfamily of subtilisin-class enzymes known as prohormone convertases. The need for this cleavage event to release bioactive products provides a crucial regulatory step for the cell. Early protein sequencing studies of various peptide hormones suggested that the dibasic sequences Lys-Lys and Lys-Arg provided the sites of cleavage (reviewed by Lazure et al., 1983). The first protease isolated in this class was the yeast kexin, which cleaves with high selectivity both synthetic peptide and protein substrates possessing Lys-Arg at the P2 and P1 sites, respectively (Fuller et al., 1989; Brenner & Fuller, 1992). Following isolation of the yeast enzyme a number of mammalian species have been cloned including furin (Van den Ouweland et al., 1990), PC1/PC3 and PC2 (Smeekens et al., 1991), and more recently the enzymes PC4, PC5, and PACE4 (Rehmtulla et al., 1993). The enzymes possess pro-domains and must therefore themselves be processed prior to activation. Maturation has been shown to occur in an autocatalytic fashion in the cases of PC2 (Matthews et al., 1994) and of furin (Creemers et al., 1993). These studies have now shown that most cleavage takes place either at Lys-Arg and Arg-Arg dibasic sites, or at an Arg-X-Lys-Arg consensus site, depending on the intracellular pathway of localization.

Mature prohormone convertases are large enzymes that typically possess 600–800 amino acids. In addition to the subtilisin-like catalytic domain, they also variously possess other structural elements such as transmembrane anchors, Ser/Thr-rich regions, glycosylation sites and Cys-rich regions (Seidah et al., 1991).

Based on homology modeling, it was predicted that these enzymes possess a greatly increased number of negatively charged residues near the substrate binding cleft. Many of these amino acids are highly conserved (Siezen et al., 1991; Fig. 5). Their importance was tested by site-directed mutagenesis of furin, using processing of a peptide hormone *in vivo* as the functional assay (Creemers et al., 1993). The following residues were mutated: Asp 33, Asp 61, Glu 101, Asp 104, Glu 107, Glu 129, Asp 130, Asp 131, Asp 165, and Asp 209. Cleavage was assayed toward the wild-type hormone precursor as well as toward three mutants in which one of the positively charged amino acids in the cleavage site sequence P4-Arg-P3-Ser-P2-Lys-P1-Arg was altered to Gly or Ala. The ability of mutants to carry out autoproteolytic activation was also assessed.

Mutation of the P1-Arg in this sequence gave rise to prohormones that could not be processed either by wild-type or by any of the mutant furins, suggesting that a basic residue at this position is critical to recognition (Creemers et al., 1993). Several of the mutants possessed preferences for one of the three mutant prohormone substrates, implicating the Asp or Glu at that enzyme position in recognition of the substrate residue that was altered. Thus, Asp 33 is implicated in P2-site binding and Glu 107 in P4-site binding, in accord with modeling that predicts their locations adjacent to these substrate positions (Siezen et al., 1991). Mutation of Asp 165, predicted to lie at the base of the S1 site, abolished activity, as did removal of the negative charge

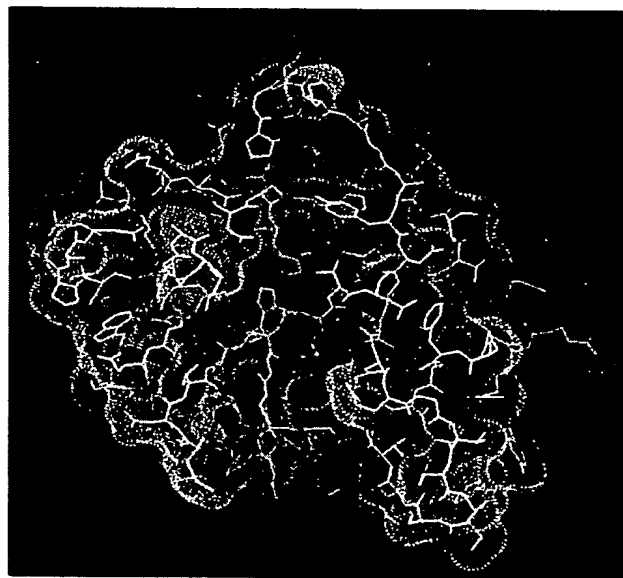
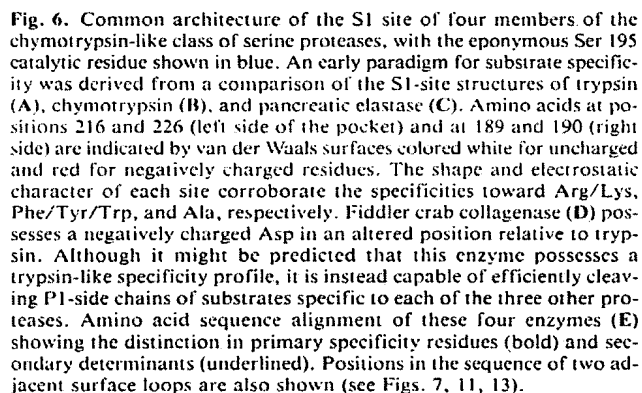


Fig. 5. A distinct subclass of the subtilisin family of serine proteases, the prohormone convertases, are involved in prohormone processing in a number of important physiological contexts. The specificity of processing is toward sites possessing 2–4 Arg and Lys residues at the P1–P4 positions. Shown is a solvent-accessible protein surface on which are mapped the binding determinants specifying prohormone processing by furin. The structure is that of subtilisin BPN' complexed to SSI because no three-dimensional structure is yet available in this subclass. A large number of negatively charged amino acids is found on the substrate binding face of the enzyme (red). The catalytic triad is in blue and the substrate is in yellow, with the P1–P4 amino acids in green.

Substrate specificity in the chymotrypsin family

Molecular modeling methods have been used to create a structure-based sequence alignment of the chymotrypsin-like serine proteases (Greer, 1990), which is very useful in assessing substrate preferences. The specificity is usually most pronounced at the S1-sites of the enzymes, where the majority of sequences group into one of three subclasses definable by inspection of a small number of crucial amino acids. Position 189, located at the base of the S1 pocket, is very highly conserved as an Asp in enzymes with trypsin-like specificity toward Arg- and Lys-containing substrates (Fig. 6; chymotrypsin numbering system is used throughout – see Greer, 1990). It is found as a Ser or other small amino acid in chymotrypsin and elastase-class enzymes, which manifest specificity toward aromatic and small hydrophobic amino acids, respectively. The amino acid side chains at positions 190 and 228 extend into the base of the pocket as well and play an additional role to modulate the specificity profile. Amino acids at positions 216 and 226 are usually Gly in both trypsin and chymotrypsin-like enzymes; larger amino acids at these positions partially or fully block access of large substrate side chains to the base of the pocket (Fig. 6). Accordingly, elastases possess larger, usually nonpolar residues at these positions.



providing a platform for interaction with small hydrophobic substrate P1-amino acids. The shapes of the S1 pockets of trypsin, chymotrypsin, and elastase thus appear to readily explain the observed specificities, leading to the canonical view that substrate preferences are in fact determined by this limited set of amino acids (Stroud, 1974). However, as discussed below, this perspective has now been shown to be incorrect by the discovery that other structural elements distant from the substrate binding site are also crucial determinants of specificity.

Kinetic measurements of substrate preferences for the two mammalian elastases of known structure (PPE and HNE) permit a more detailed appraisal of structure-function relationships

(Bode et al., 1989b). Both enzymes possess bowl-shaped hydrophobic S1 binding sites that accommodate small hydrophobic substrates (Watson et al., 1970; Navia et al., 1989). However, the S1 site of PPE has been described as slightly less hydrophobic and marginally smaller than that of HNE (Bode et al., 1989b). PPE cleaves peptide bonds preferentially at small P1-Ala and Nva side chains (Harper et al., 1984), whereas HNE manifests substantial activity toward the branched-chain Val, Ile, and Leu residues (Harper et al., 1984; Stein et al., 1987). These preferences are in accord with the smaller S1 site of PPE, but the small difference in size is insufficient to account for the altered profiles. The identity of the amino acids that line the S1 pockets differ substantially in the two enzymes, most notably by the presence of the charged Asp 226 in HNE, which is present as a Thr in PPE. In HNE, Asp 226 is buried by Val 216 and Val 190, and the carboxylate group points away from substrate into a network of buried water molecules (Navia et al., 1989). One possible explanation for the superior ability of HNE to cleave branched-chain substrates could thus be that the S1-site possesses greater intrinsic flexibility as a consequence of its different construction and interaction with surrounding portions of the structure (Bode et al., 1989b). A small shrinkage of the S1 site is in fact observed upon binding Val relative to Leu in this position (Bode et al., 1986b; Wei et al., 1988).

Cleavage of peptide substrates adjacent to the acidic Asp and Glu residues is the hallmark of an additional subclass of enzymes. Recognition of the negatively charged carboxylate is accomplished by means of a His residue at position 213 in a number of microbial enzymes including the *Staphylococcus aureus* V8 protease (Drapeau, 1978), SGPE (Svendsen et al., 1991), and two epidermolytic toxins of *S. aureus* (Dancer et al., 1990). Recently, the crystal structure of SGPE complexed with the tetrapeptide Ala-Ala-Pro-Glu has been determined at 2.0 Å resolution (Nienaber et al., 1993). The structure reveals that the Glu carboxylate is indeed bound directly by His 213 as well as by the side chains of Ser 192 and Ser 216. The structure of the enzyme also shows that His 213 is hydrogen bonded in series to two other His residues at positions 199 and 228 to form a solvent-inaccessible His triad that penetrates through the core of the enzyme. This remarkable structural feature is postulated to play a role in substrate charge compensation, by delocalizing the substrate negative charge through proton transfer across the His residues (Nienaber et al., 1993). No other serine protease is known to possess the His triad. An alternative to the use of His 213 is found in a protease from cytotoxic T-lymphocytes, which possesses an Arg at position 226 (Murphy et al., 1988). This enzyme is unusual in its preference for cleavage at Asp rather than Glu residues (Otake et al., 1991). Mutation of Arg 226 to Gly, followed by qualitative assay of crude lysates in which the variant was expressed, showed lowered activity toward peptidyl P1-Asp thio-benzyl ester substrates and increased activity toward analogous P1-Phe substrates (Caputo et al., 1994).

Virtually all chymotrypsin-like serine proteases share a common feature: an S1-site specificity that is restricted to a relatively narrow subset of the naturally occurring amino acids. It therefore came as some surprise when one enzyme, the collagenolytic serine protease I from the fiddler crab *Uca pugilator*, was shown to possess high catalytic activity toward each of trypsin, chymotrypsin, and elastase-like substrates (Grant & Eisen, 1980). The specificity profile of this enzyme has recently been reexamined in detail (Tsu et al., 1994). Crab collagenase exhibits 5% of clas-

tase, 10% of chymotrypsin, and 65% of trypsin activity, as assessed by k_{cat}/K_m values toward peptidyl amide substrates possessing Ala, Phe, and Arg, respectively, at the P1 position. k_{cat} values toward each of these amino acids are extremely high. Additionally, it is the most efficient chymotrypsin-like enzyme known toward P1-Leu and P1-Gln amide substrates, manifesting 6-fold and 50-fold greater activities than does chymotrypsin toward these substrates (Tsu et al., 1994). Therefore, the chymotrypsin-like scaffold can maintain an S1 binding pocket that accommodates a very broad range of amino acids without sacrificing catalytic efficiency.

Crab collagenase exhibits an interesting rearrangement of a negative charge at the base of the S1 site: residues Asp 189 and Gly 226 of trypsin are altered to Gly 189 and Asp 226 in collagenase (Grant et al., 1980; Fig. 6). However, this predicts a strict specificity for P1-Lys and Arg substrates: the amino acids at positions 190 and 216 are Thr and Gly, respectively, which allows access of the substrate to Asp 226. As discussed above, Asp 226 of human neutrophil elastase is buried by Val 216, leading to a hydrophobic specificity profile (Navia et al., 1989). A possible explanation for the ability of crab collagenase to accommodate hydrophobic as well as positively charged substrate residues is provided by a recently refined 2.5-Å crystal structure of the enzyme complexed with the dimeric serine protease inhibitor ecotin (J.J. Perona, C.A. Tsu, C.S. Craik, & R.J. Fletterick, submitted for publication). The structure shows that one carboxylate oxygen of Asp 226 is accessible to substrate, but that the P1-methionine residue of ecotin does not enter the S1-site and binds instead on the surface of the enzyme adjacent to the disulfide bond at positions 191–220. Modeling shows that the pocket can provide multiple binding sites that accommodate diverse amino acid side chains in distinct positions. Therefore, S1-site flexibility does not appear to be utilized as a structural determinant in the broad specificity of crab collagenase.

α -Lytic protease: Exploring the role of structural plasticity in substrate specificity

α -Lytic protease, an extracellular enzyme produced by the soil bacterium *L. enzymogenes*, has been the subject of intensive analysis aimed at relating structure to catalytic activity. This microbial protease, while possessing the chymotrypsin-like fold comprising two β -barrels (Brayer et al., 1979), nevertheless displays large insertions and deletions relative to the pancreatic enzymes, resulting in an overall RMS deviation in the positions of structurally equivalent α -carbons of 1.36 Å for 110 of 198 amino acids, when compared with chymotrypsin (Fujinaga et al., 1985). By comparison, the equivalent pairwise fits with the bacterial proteases SGPA and SGPB yield RMS deviations of roughly 0.7 Å, a value very similar to that which relates the mammalian pancreatic enzymes to each other. The S1 pockets of α -lytic protease and trypsin are particularly divergent in structure (Fig. 7). An insertion of two amino acids causes Met 192 of α -lytic protease to occupy a position similar to Ser 190 of trypsin. More strikingly, an adjacent surface loop at positions 185–188 is deleted in α -lytic protease, and a second nearby loop at positions 217–225 is enlarged by eight amino acids. A consequence of these differences is that, although both enzymes possess a disulfide bond linking the conserved residues Cys 191 and Cys 220, the positions of the sulfur atoms are displaced by 7–8 Å (Fig. 7).

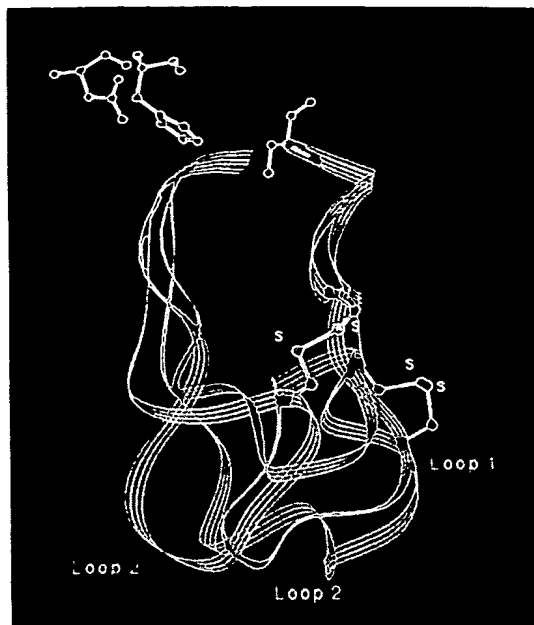


Fig. 7. Diversity in S1-site structure between the mammalian and the microbial trypsin-like enzymes is illustrated by a superposition of trypsin (green) and α -lytic protease (red). Although the mammalian enzymes such as trypsin possess two well-defined loops (loop 1 and loop 2) joining the β -strands of the specificity pocket, in α -lytic protease and other microbial enzymes loop 1 is absent, whereas loop 2 is greatly enlarged. Conserved disulfide bonds of each enzyme (Cys 191-Cys 220; yellow) are displaced some 7 Å from each other. The catalytic triad is shown at the top in green.

Kinetic data show that α -lytic protease possesses a hydrophobic specificity profile for substrate residues in the P1 position. The preference of the enzyme at P1, as described by relative k_{cat}/K_m values, is roughly Ala > Met, Val, Gly > Nle > Leu > Phe for hydrolysis of tetrapeptide amide substrates (Bauer et al., 1981; Bone et al., 1991). The structural elements that interact with the P1-substrate side chains comprise the three hydrophobic side chains Met 192, Met 213, and Val 217a, which together form a shallow depression in the enzyme surface (Brayer et al., 1979; Fujinaga et al., 1985; Fig. 8). More recently, six crystal structures of the enzyme complexed with peptidyl boronic acid inhibitors of the general structure R-boroX (where R is methoxysuccinyl-Ala-Ala-Pro and boroX is the α -aminoboronic acid analog of Ala, Val, Ile, Nle, Leu, or Phe) have been determined at resolutions between 2.0 and 2.5 Å (Bone et al., 1987, 1989a, 1991). Boronic acids are tight-binding (K_i 's in the nanomolar range) reversible inhibitors of serine proteases (Kettner & Shenvi, 1984) that form covalent, nearly tetrahedral adducts with Ser 195 (Bone et al., 1987). They represent good structural analogs of the high-energy tetrahedral intermediate present on the actual catalytic pathway.

The crystal structures of the boronic acid complexes confirm that covalent tetrahedral adducts are formed with O γ of Ser 195 for the P1-Ala, Val, Ile, Leu, and Nle inhibitors. The large P1-Phe side chain cannot fit into the S1-site, leading to the formation of an unusual trigonal adduct that includes His 57 (Bone

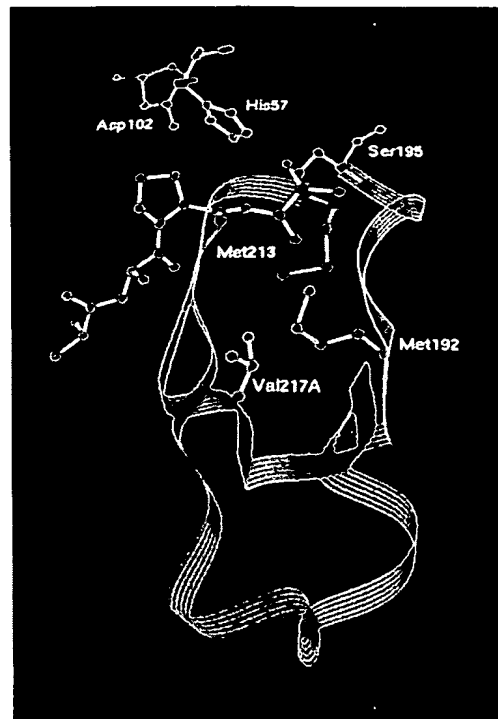


Fig. 8. Structure of the S1 site of α -lytic protease bound to the substrate analog *suc*-Ala-Ala-Pro-Ala-boronic acid (red), showing the positions of the hydrophobic amino acids Met 192, Met 213, and Val 217a, which form a platform for binding of small hydrophobic side chains. The three β -strands of the S1 site are shown in yellow and the large connecting ω -loop is in green. Catalytic groups are also in green (top). Mutation of either Met 192 or Met 213 to Ala creates variant enzymes possessing greatly broadened specificities toward hydrophobic amino acids, without sacrificing catalytic efficiency.

et al., 1989a). The interactions of the inhibitor among these structures are nearly identical with the exception of the way in which the P1 side-chains interact with Met 192, Met 213, and Val 217a. These side chains adjust conformation in response to the differing sizes and shapes of the inhibitor amino acids. Small shifts in the position of adjacent main-chain atoms in the S1 and S2 specificity sites occur in the complexes with the larger Nle and Phe. Particular importance has been ascribed to the rearrangements at positions 217a-217d (Bone et al., 1989a, 1991; see below). Low activity toward the larger Leu and Phe side chains appears to arise solely from steric considerations, whereas Met is preferred to Leu presumably owing to its greater flexibility. Although the structural basis for the preference of Ala relative to Val was not unambiguously clear, it was proposed that strong binding to the oxyanion hole, required in the transition state, is prevented for the Val substrate on steric grounds. Differences in the electronic character of the boronate inhibitor, relative to a true transition state, do not allow for a complete mimicking of the latter (Bone et al., 1989a).

The substrate specificity profile of α -lytic protease was altered dramatically by the introduction of either of two single-site mutations in the S1 site: M192A or M213A (Bone et al., 1989b; Ta-

ble 2; Figs. 8, 9). In each case, high activity toward Ala was retained, but the increased size of the S1 pocket allowed accommodation of P1-side chains as large as Phe, with catalytic efficiencies k_{cat}/K_m increased up to 15-fold relative to wild-type cleavage at P1-Ala. For M192A, improved catalytic efficiencies toward P1-Met and P1-Val resulted mainly from lowered K_m values, whereas the P1-Leu and P1-Phe substrates were improved in both k_{cat} and K_m . The catalytic activity toward P1-Leu and P1-Phe substrates was improved by 10^4 – 10^6 -fold, respectively, relative to wild type. However, the wild-type preference of nearly 10^5 -fold for P1-Ala/Phe was decreased to 30-fold in M192A and nearly completely eliminated in M213A (Table 2). Complicating a straightforward interpretation of the profiles of these variants were two factors: (1) the dependence of k_{cat} , K_m , and k_{cat}/K_m was not correlated with the size or hydrophobicity of the P1 side-chain; (2) enlargement of the pocket by the same volume in the two mutants gave rise to considerably different functional effects. Therefore, extensive structural analysis of the mutant enzymes complexed with the boronic acid inhibitors was carried out to understand which factors cause the altered specificities (Bone et al., 1989b, 1991).

The principle rationale for the exceptionally broad specificity profiles of M192A and M213A is that the S1 site possesses structural plasticity, which encompasses a combination of alternate side-chain conformations as well as deformability of the main chain (Bone et al., 1989b; Fig. 9). For example, accommodation of the P1-Phe side chain by M192A results from a substrate-induced conformational change, in which the side chain of Val 217a rotates to remove one carbon from the pocket, and the main chain from Val 217a to Val 217d shifts by 0.5–0.8 Å. This permits the large inflexible aromatic ring to be nearly completely buried in the specificity pocket. In this case, some of the binding energy is presumably used to drive the conformational change in the protein, a phenomenon that is also observed to lesser extents in other mutant-inhibitor complexes. In general, hydrogen bond lengths, buried hydrophobic surface area, unfilled cavity volume, and the magnitude of conformational changes vary significantly among the various mutant and wild-type complexes (Bone et al., 1991). The energetic consequences of these differences were quantified (see Bone & Agard [1991] for a review of the energetics of intermolecular interactions) and correlated with free energies of catalysis for the various mutant-substrate combinations.

The analysis has led to an increased understanding of the way in which the different energetic terms can contribute to the stabilization of the enzyme-substrate complex, although no single factor has been found that consistently correlates well with ei-

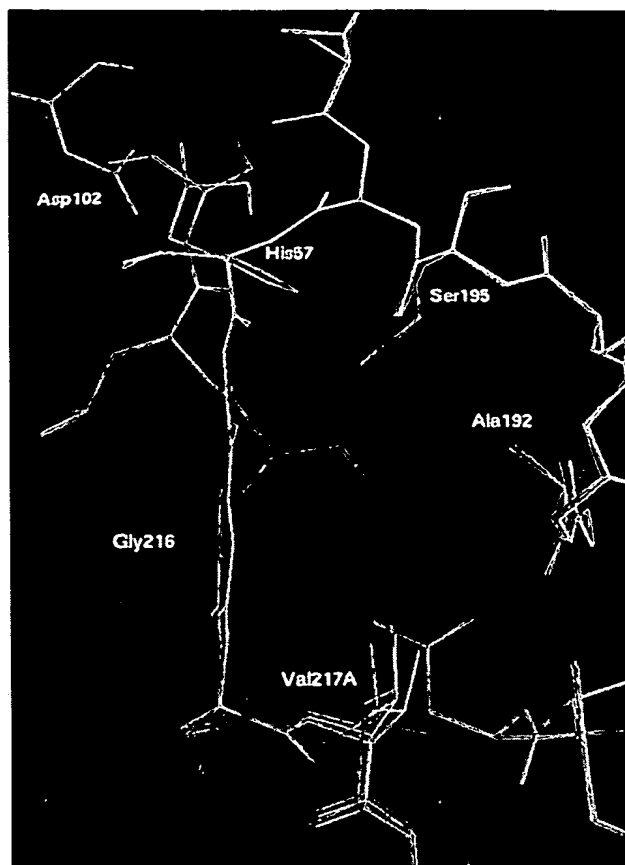


Fig. 9. Principal rationale for the ability of α -lytic protease mutants to exhibit greatly enhanced specificities toward new substrate side chains is structural plasticity of the S1 site. Shown is a superposition of five structures of the M192A variant of the enzyme (the new Ala 192 side chain is at the right side). Each enzyme is complexed with a peptidyl boronate inhibitor (not shown for clarity) possessing a particular hydrophobic P1-side chain (see Fig. 8 for inhibitor binding). The conformation of the active site adjusts to the different substrates at position Gly 216 and in the following loop region (bottom). Both side-chain and main-chain rearrangements are important components of active-site plasticity. The ability of the active site to adjust in this manner may be an important factor in the ability to effect specificity modification by mutation at only a single site.

Table 2. Broadening the specificity of α -lytic protease^a

X	Wild type	M192A	M213A
Ala	21,000	10,000	600
Val	790	3,000	340
Met	1,800	35,000	980
Leu	4.1	11,000	160
Phe	0.38	31,000	340

^a Substrate: *suc*-Ala-Ala-Pro-X-pNA. k_{cat}/K_m , s⁻¹ M⁻¹.

ther activity or inhibition (Bone et al., 1991). Thus, the wild-type enzyme has a relatively limited ability to adapt to large side chains, so that the specificity profile is driven primarily by steric exclusion. M192A, however, is improved in its ability to hydrolyze large side chains in part because the degree of conformational change required for their accommodation is reduced; further, it also possesses the ability to shrink so that P1-Ala substrates are hydrolyzed well. By contrast, the M213A pocket cannot contract, leading to a sharply reduced activity toward P1-Ala as well as a reduced discrimination relative to P1-Gly (Bone et al., 1991). In both mutants, however, the broad specificities depend on the ability of the main chain and side chain atoms at positions 217a–217d to readjust (Fig. 9). This flexibility is proposed to arise from a large adjacent surface loop, which begins at res-

idue 217a (Figs. 7, 8), and which appears to be able to absorb structural changes in the preceding residues. The energies of interaction of the S1 site with this and other peripheral structural elements thus also play a significant role in determining the specificity profiles.

Another recent study of α -lytic protease used random mutagenesis of four residues in the substrate binding pocket, coupled to an activity screen using synthetic substrates, to identify new variants with altered specificities (Graham et al., 1993). A library was constructed beginning with the M192A variant, with randomization of positions Gly 192a, Arg 192b, Met 213, and Val 217a. Screening and qualitative characterization of 47 active variants revealed that a majority of the enzymes retained a specificity profile similar to that of the parent M192A. Also emerging from the screen was a subclass of enzymes capable of cleaving P1-His-containing substrates. All mutants possessing this ability contained His 213, an amino acid heretofore correlated with P1-Glu specificity in other microbial enzymes (Nienaber et al., 1993). In general, residue 213 appears to play a significant role as a primary specificity determinant in several microbial enzymes. Although this amino acid has not yet been mutated in any mammalian protease, it appears very unlikely that it will assume a similar role. Clearly the divergence in structure of the S1 site in the two subclasses (Fig. 7) has led to a more prominent role for this residue in the bacterial enzymes, despite the fact that its position relative to the Ser 195/His 57 catalytic couple does not vary.

Kinetic data indicate that α -lytic protease makes substrate binding interactions over at least six subsites from P2' to P4 (Bauer et al., 1981). Interestingly, the crystal structure shows that a small hydrophobic pocket exists beyond the P4 side chain of the tetrapeptide boronic acid inhibitor, formed from residues Leu 227, Leu 180, Val 167, Ala 169, and Ser 225 (Bone et al., 1987). Although extension of a substrate side chain to fill the S5 site does not have a significant influence on kinetic parameters (Bauer et al., 1981), it is possible that additional binding energy from interactions in the hydrophobic pocket cannot be realized in catalysis unless a P6 side-chain is also bound. Little specificity has been observed at the other subsites, although a preference for Pro at position P2 has been noted in binding of the peptide boronic acid inhibitors (Bone et al., 1987). Although the S2 enzyme site is hydrophobic, adjacent side-chain hydroxyl groups of Ser 214 and Tyr 171 participate in a hydrogen bonding network, which includes the carboxylate of Asp 102. Introduction of the mutations S214A and Y171F caused decreases in both k_{cat} and K_m , and the data were used to infer that the role of the two hydroxyl groups in the native enzyme is to facilitate catalysis by maintaining the S2 site in an optimal configuration (Epstein & Abeles, 1992).

Mutational analysis of trypsin: Combining structural genetics, classical enzymology, and X-ray crystallography

Trypsin represents the third serine protease that has been the subject of extensive mutational analysis aimed at an understanding of substrate specificity. These studies have focused largely on the origins of specificity at the primary S1 site. At this position, trypsin hydrolyzes amide substrates containing P1-Lys and P1-Arg amino acids by factors of 10^5 or greater relative to the next-preferred residues (Graf et al., 1988; Evnin et al., 1990).

The preference of the enzyme is 2-10-fold in favor of Arg- relative to Lys-containing substrates (Craik et al., 1985; Perona et al., 1993c). As might be expected from their structural disparity, Lys and Arg interact in a differential manner with the primary determinants Asp 189 and Ser 190 (Ruhlmann et al., 1973; Bode et al., 1984; Fig. 10). The guanidinium group of P1-Arg substrates makes an ion-pair interaction with Asp 189, whereas the interaction of P1-Lys is solely by a water-mediated contact. Both Arg and Lys substrate side chains also interact with Ser 190.

An early study assessed the precision with which the S1 site is constructed by introducing small perturbations: the Gly residues at positions 216 and 226 were converted to Ala, resulting in the three trypsin mutants G216A, G226A and G216A/G226A (Craik et al., 1985; Fig. 10). Relative specificities for tripeptide amide P1-Arg/Lys substrates, as assessed by the ratio of k_{cat}/K_m values, were altered by up to 20-fold. Catalytic efficiencies were decreased by 40-fold to 10^4 -fold, and these effects involved significant decreases in k_{cat} as well as higher K_m values. The differential effects of the k_{cat} and K_m values resulted in enzymes that were more Arg specific (G216A) and more Lys specific (G226A) than the wild-type enzyme. Subsequent crystal structure determinations of trypsins G226A (Wilke et al., 1991) and G216A (M.E. McGrath & R.J. Fletcher, unpubl. results)

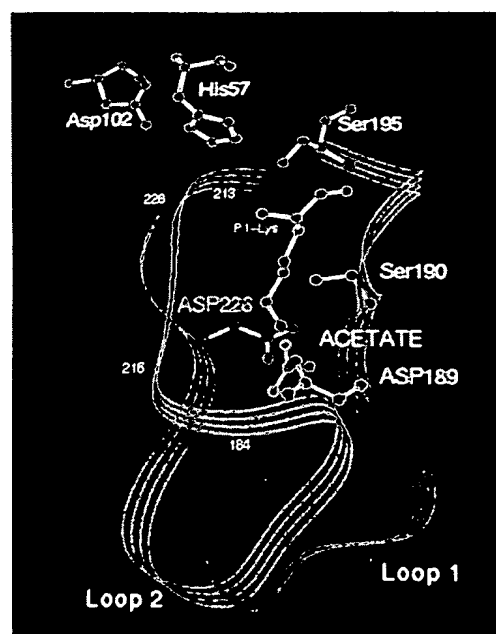


Fig. 10. Role of the position of the negative charge at the base of the trypsin S1 site has been probed by random and site-directed mutagenesis coupled to crystal structure analysis of variants. Shown is the structure of the S1 binding pocket of trypsin, indicating the positions at which the negatively charged amino acid has been determined by X-ray crystal structures. Blue, wild-type trypsin at position 189; red, trypsin D189G/G226D at position 226; yellow, exogenously added acetate ion in trypsin D189S (acetate reconstitutes activity toward P1-Arg and P1-Lys-containing substrates). Wild-type amino acids at positions 216 and 226 are each Gly, permitting access of the large P1-Lys (green) and P1-Arg side chains to Asp 189.

complexed with benzamidine showed that the alanine substitutions produced no structural perturbations beyond the immediate vicinity of the mutated residues. Because the catalytic triad Ser 195, His 57, and Asp 102 amino acids are unaffected by these binding pocket alterations, it is highly probable that the decreases in k_{cat} are attributable to altering the catalytic register of the scissile bond. These data thus provided an early demonstration that substrate binding and catalytic turnover are interrelated functions in trypsin, and that they can be affected differentially to alter the function of the enzyme.

A series of studies have addressed the role of the negatively charged Asp 189 residue in binding and catalysis. These investigations have made use of both site-directed mutagenesis as well as a genetic selection approach for the isolation of new variants (Fig. 4B). The selection is based on expression of a library of trypsin variants into the periplasmic space of an *E. coli* strain that is auxotrophic for arginine or lysine (Evnin et al., 1990). Cells are plated on minimal media containing a nonnutritive substrate analog of one of these amino acids; active trypsins cleave the analog, liberating free amino acid and thereby relieving the auxotrophy (Evnin et al., 1990; Perona et al., 1993a).

Twenty variant trypsins have been isolated from a library of 400 possible mutants encompassing the amino acids at positions 189 and 190 at the base of the S1 site. Kinetic characterization of these enzymes, as well as of the variants D189K (Graf et al., 1987) and D189S (Graf et al., 1988), indicates that the presence of a negative charge at the base of the binding pocket is essential to high-level catalysis by trypsin. Variants lacking the negative charge are compromised in k_{cat}/K_m toward peptidyl Arg- or Lys-containing amide substrates by a factor of 10^5 or greater. Activity toward these substrates is partially restored by the presence of an Asp or Glu residue at positions 189 or 190. The variants span a range of catalytic efficiencies ranging from wild type to decreases of 10^6 -fold (Evnin et al., 1990; Perona et al., 1993a).

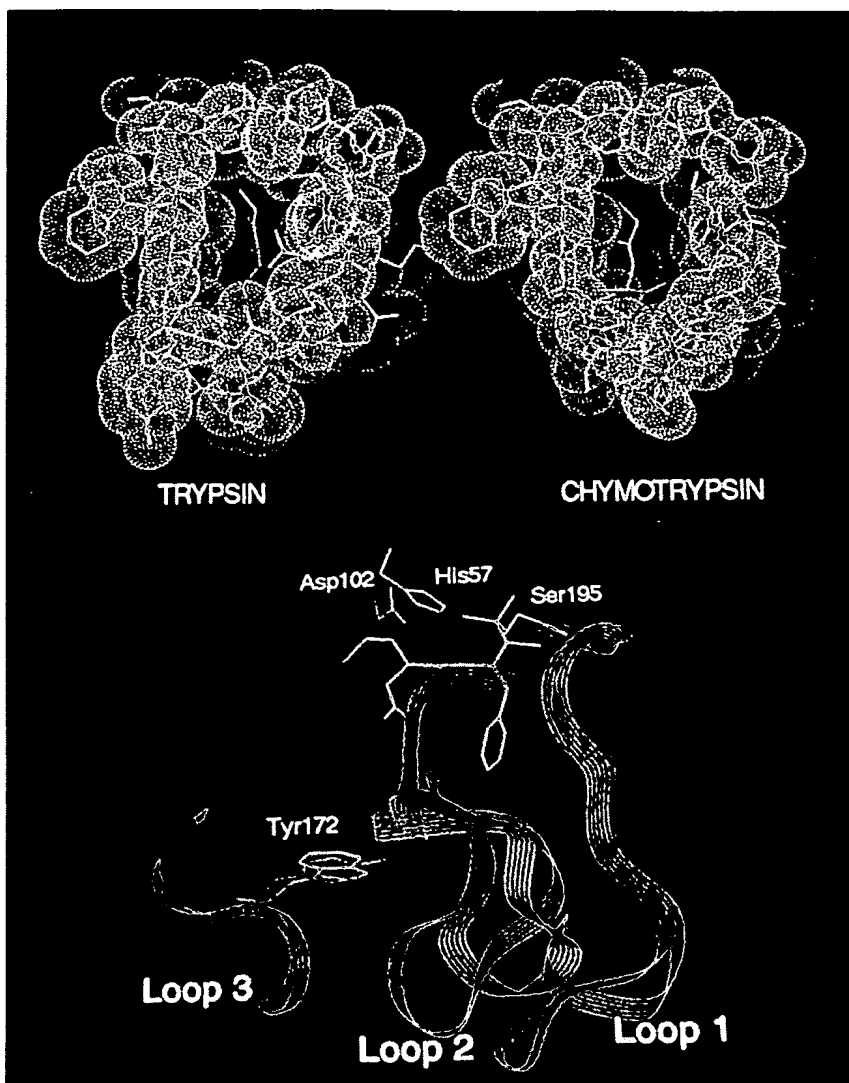
A framework for the interpretation of these data is provided by kinetic and crystallographic investigation of two other variants: trypsins D189G/G226D (Perona et al., 1993b, 1993c) and D189S (Perona et al., 1994). The structure of each mutant enzyme was determined complexed with the protein inhibitors APPI and/or BPTI, which are analogs of the substrate Michaelis complexes possessing Arg and Lys, respectively, at the P1 position (Perona et al., 1993b). This allows for the direct comparison of substrate-like interactions of Arg and Lys side chains in the binding pockets of wild-type and mutant enzymes. Trypsin D189G/G226D is equally reduced (10-fold) in binding affinity toward Lys and Arg substrates and is sharply lowered (10^3 -fold) in k_{cat} toward Arg. The crystallographic analysis showed that Asp 226 is partially sequestered from substrate by intramolecular interactions made with Ser 190 and Tyr 228, such that only a single carboxylate oxygen is available for substrate binding. Further, comparisons with the wild-type interactions indicated no correlation between the binding affinities of either Lys and Arg substrates and the number of direct contacts made with Asp 226. Therefore, it appears that substrate binding affinity to trypsin depends upon the accessibility of the negative charge to substrate and not upon the formation of direct interactions. This observation implies that direct electrostatic hydrogen bonding interactions between the substrate Lys/Arg and the enzyme carboxylate group do not significantly improve the free energy of binding relative to indirect water-mediated interactions (Perona et al., 1993c).

The crystal structure of trypsin D189S revealed that an acetate ion from the crystallization buffer was trapped at the base of the binding pocket, such that its carboxylate group was partially oriented toward substrate (Perona et al., 1994; Fig. 10). Exogenously added acetate provided up to 300-fold rate enhancements to trypsin D189S toward Arg- and Lys-containing substrates, but catalytic activity remained diminished relative to wild-type trypsin. This structure thus provides a second example showing that optimal placement of the negative charge in the binding pocket is critical to catalysis. Significantly, the diminished activities of both trypsins D189G/G226D and D189S/acetate are reflected in k_{cat} as well as K_m . Measurement of activities toward analogous ester as well as amide substrates by these enzymes allows calculation of the mechanistic parameters K_1 , k_2 , and k_3 (Zerner & Bender, 1964; Fig. 2C), removing the ambiguity in interpretation of the steady-state Michaelis-Menten parameters. This analysis shows that the role of the Asp 189 carboxylate in trypsin is twofold: it provides both tight binding affinity K_1 as well as high acylation rate k_2 (Perona et al., 1994). Therefore, the precise location of the negatively charged group within the trypsin S1 site is critical to positioning the scissile bond in catalytic register with Ser 195 and His 57.

Analysis of the kinetic properties of the 20 variants isolated from the genetic selection corroborates these hypotheses regarding the operation of the S1 site. Although the binding constants of the enzymes vary widely, it is significant that relative affinities for Lys versus Arg substrates remain very similar (Perona et al., 1993a). The negatively charged carboxylate in these mutants is provided by either Asp or Glu at positions 189 or 190, and the partner to this residue is 1 of 10 different amino acids. Thus, it is very unlikely that equal reductions in affinity toward Lys versus Arg substrates can in most cases be attributed to an equal loss of hydrogen bonding or electrostatic interactions. Instead, binding affinity is likely to be better correlated with accessibility of the negative charge to substrate; barring substrate-induced conformational changes, this accessibility will be the same for both Lys and Arg substrates. Binding affinities are then predicted to be weaker when the carboxylate is partially sequestered from substrates, as seen in the structures of the mutants D189G/G226D and D189S/acetate. Crystal structures of additional variants from the selection pool should enable a quantitative correlation between binding affinity and accessibility of the negative charge. These experiments also explain the rationale for conservation of the Asp at position 189 in the vast majority of trypsin homologs, because other locations result in partial sequestration of the negative charge.

In a second set of experiments, site-directed mutagenesis has been used to convert trypsin into a chymotrypsin-like protease possessing high selectivity for cleavage adjacent to large hydrophobic amino acids (Hedstrom et al., 1992, 1994a, 1994b). The structures of the S1 pockets of the two enzymes are very similar (Figs. 6, 11A), so it was expected that specificity modification might be straightforward as in subtilisin and α -lytic protease. However, when the amino acids directly in contact with substrate were exchanged into trypsin, the resulting variants D189S and D189S/Q192M/I138T/T218 failed to exhibit significant improvement in cleavage of P1-Phe amide substrates (Graf et al., 1988; Hedstrom et al., 1992; Table 3). Poor efficiency was also shown toward trypsin substrates, as expected because the pocket lacks a negative charge. The crystal structure of trypsin D189S showed that only very local structural changes were introduced

A



B

Fig. 11. A: Comparison of the S1 sites of trypsin and chymotrypsin. Van der Waals surfaces of each enzyme are shown with the position-189 amino acid (Asp in trypsin; Ser in chymotrypsin) indicated in red. In yellow is the conserved Ser 190, which is oriented into the S1 pocket in trypsin but rotates out in chymotrypsin. The inserted Thr 218 in chymotrypsin is shown in green. Two other amino acids directly in or adjacent to the S1 site are Ile 138 (Thr 138 in chymotrypsin), and Gln 192 (Met 192 in chymotrypsin). Although a high degree of structural similarity is clear, exchange of these four amino acids fails to transfer chymotryptic specificity to trypsin. **B:** Structural determinants required to exchange substrate specificity include two adjacent surface loops (loop 1 and loop 2) and an amino acid (Tyr 172 in trypsin) in a third adjacent segment (loop 3). None of these structural elements directly contact substrate (shown at top in thin green lines). Trypsin is shown in red and chymotrypsin in green.

as a consequence of the substitution; the binding pocket maintains a trypsin-like conformation (Perona et al., 1994). This confirms that the small structural differences between trypsin and chymotrypsin in the S1 site (Fig. 11A) must be critical determinants of the specificity and must rely on more distant parts of the structure for maintenance of their particular conformations.

Exchange of the two surface loops, loop 1 and loop 2 (Fig. 11B), resulted in the hybrid enzyme Tr→Ch[S1+L1+L2], which exhibited an acylation rate constant k_2 equal to that of chymotrypsin toward peptidyl P1-Phe amide substrates (Hedstrom et al., 1992; Table 3). However, the enzyme was still reduced by nearly 10^3 -fold in k_{cat}/K_m because of a very weak substrate binding affinity. The mechanistic kinetic parameters K_1 , k_2 , and k_3 were calculated for cleavage of both single-residue and peptidyl P1-Phe amide substrates for the enzymes trypsin, chymotrypsin, D189S and Tr→Ch[S1+L1+L2]. These data showed that, like chymotrypsin, the hybrid trypsin was able to use the

binding energy obtained by occupancy of the S2-S4 enzyme sites to increase the acylation rate. They also demonstrated that, among this series of enzymes, the key mechanistic step that determines substrate specificity is not binding affinity, but instead the chemical step of acylation (Hedstrom et al., 1992, 1994a).

Further mutations were sought to improve catalytic efficiency toward chymotryptic substrates by increasing binding affinity. The additional mutation Y172W in a third adjacent surface loop (Fig. 11B) produced the hybrid enzyme Tr→Ch[S1+L1+L2+Y172W], which improves the activity of Tr→Ch[S1+L1+L2] by 20–50-fold, creating an enzyme with up to 15% of the activity of chymotrypsin (Hedstrom et al., 1994b; Table 3). The improvement toward a tetrapeptide P1-Phe amide substrate is manifested almost entirely in tighter binding affinity. The relative catalytic efficiencies measured toward Trp, Tyr, Phe, and Leu P1-amide substrates also more closely mimic chymotrypsin (Hedstrom et al., 1994b).

Table 3. Conversion of trypsin to chymotryptic specificity^a

	K_i (M)	k_2 (s ⁻¹)	k_3 (s ⁻¹)
Trypsin	>0.25	>0.2	36
D189S	0.015	0.29	33
Tr→Ch[S1+L1+L2]	0.011	20	37
Tr→Ch[S1+L1+L2+Y172W]	5.0×10^{-4}	41	63
Chymotrypsin	1.5×10^{-3}	850	52

^a Substrate: *suc*-Ala-Ala-Pro-Phe-pNA.

The structural basis for the activities of the two hybrid trypsin was elucidated by determination of their crystal structures complexed with the transition-state inactivator *suc*-Ala-Ala-Pro-Phe-chloromethyl ketone (*suc*-AAPF-CMK; Perona et al., 1995). Loop 2 of Tr→Ch[S1+L1+L2] adopts a conformation identical to that which it possesses in chymotrypsin. However, amino acids at positions 185–187 within Loop 1 are disordered. The structure of Tr→Ch[S1+L1+L2+Y172W] showed improved order in Loop 1 and a rearrangement of solvent structure and Ser 217 side-chain orientation, each of which more closely mimicked the structure of chymotrypsin. No other changes were present between the two hybrid enzymes, implicating these structural elements as important determinants of K_i in chymotrypsin.

Both hybrid enzymes possess wild-type chymotrypsin-like acylation rates k_2 toward peptidyl P1-Phe amide substrates, and each utilizes binding of the extended peptide (substrate sites P2–P4) to increase this rate. In fact, the 10⁶-fold specificity of chymotrypsin relative to trypsin for cleavage at P1-Phe is manifested solely in extended peptidyl substrates; only a 10²-fold level of discrimination exists for single-residue substrates (Hedstrom et al., 1994b). In all available crystal structures of the enzymes, including those of the trypsin hybrids, two hydrogen bonds are formed in an antiparallel β -sheet fashion with the backbone amide group of Gly 216 (Perona et al., 1995). The backbone conformation at Gly 216 differs between trypsin and chymotrypsin; the hybrid enzymes adopt a chymotrypsin-like conformation. This suggests that the Gly 216 backbone is a critical specificity determinant because it directly binds a portion of substrate responsible for a 10²-fold preference at position P1. The mechanism by which Gly 216 functions is likely to be through promoting accurate scissile bond positioning (Perona et al., 1995). Because Asp 189 of trypsin also plays a critical role in this function, it appears that the identity of the amino acid at position 189, and the backbone conformation at Gly 216, must be matched in order to permit efficient and specific catalysis by trypsin and chymotrypsin.

Structural comparisons among a number of the chymotrypsin-like proteases, including both PPE and HNE, showed a striking correlation between the P1-site specificity and the backbone conformation at position 216 (Perona et al., 1995). Three structural classes were delineated, which correspond to trypsin, chymotrypsin, and elastase-like enzymes (Fig. 12). The role of Gly 216 in promoting accurate substrate positioning may thus be a feature of many enzymes in the family. In this context it is relevant to note that the kinetic phenomenon observed for both trypsin (Perona et al., 1993c) and chymotrypsin (Hedstrom et al., 1992), namely that subsite occupancy causes large increases in the rates of the chemical steps of catalysis, is also common to other trypsin-like enzymes including PPE (Thompson & Blout,

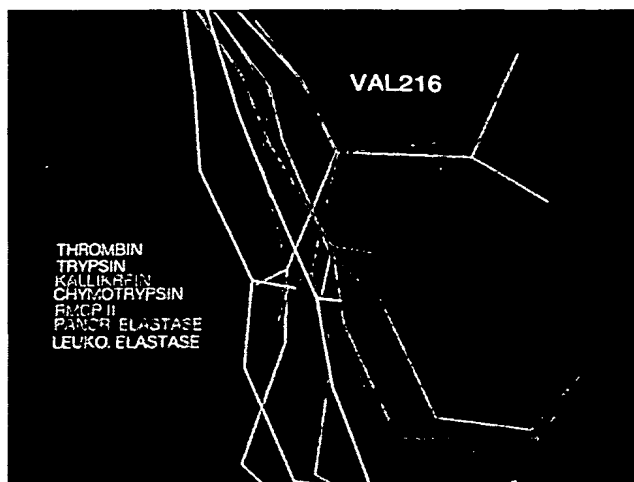


Fig. 12. A correlation is observed between the backbone conformation of residue 216 and the S1 site substrate preference among all of the trypsin-, chymotrypsin-, and elastase-like proteases of known structure. Shown is a superposition of seven mammalian serine proteases (color-coded), indicating the structure at this position that is most easily visualized in the orientation of the carbonyl oxygen atom. Specific trypsin-like, chymotrypsin-like, and elastase-like ϕ - ψ backbone angles are observed. Residue 216 binds the P3 position of the substrate in all the enzymes. Extended peptide binding to residue 216 is required both to achieve full catalytic potency as well as to obtain a maximal level of P1-site discrimination among alternative amino acids. Conversion of the substrate specificity of trypsin to that of chymotrypsin requires reorientation of Gly 216 to a chymotrypsin-like conformation. Thus, the position-216 backbone is strongly suggested as an essential specificity determinant in the mammalian trypsin-like proteases.

1970), HNE (Stein et al., 1987), SGPA (Bauer et al., 1976; Bauer, 1978), SGPB (Bauer, 1978), and α -lytic protease (Bauer et al., 1981; also see above). The significance of the recent kinetic analysis (Hedstrom et al., 1992) is that it shows that both the catalytic rate toward cognate substrates, as well as the degree of specificity at the P1-position, are dependent on the filling of subsites, which themselves exhibit little amino acid preference.

The crystal structures of the trypsin hybrids also address another fundamental question in enzyme catalysis: the role of the global protein structure. Distal structural elements such as Trp 172 and loops 1 and 2 play a key role in specifying the conformation of residues that do interact directly with substrate. Thus, their role is not solely to provide an inert platform that stabilizes the amino acids that interact directly with substrate. These elements of the global architecture play an active role in determining substrate specificity as well, which should thus be viewed as a more distributed property of the protein fold. An alternative mechanism for the way in which global protein folds may influence specificity is by modulating the degree of backbone flexibility of the S1 site, as exemplified in the α -lytic protease studies (Bone et al., 1991).

Exchange of the S1-site residues of HNE into trypsin also fails to convert the specificity of trypsin and results, as in the case of the mutants D189S and D189S/Q192M/I138T/T218, in a poor nonspecific protease (J.J. Perona & C.S. Craik, unpubl.

obs.). Similarly, introduction of Lys, Arg, or His residues into the trypsin S1 site has failed to generate specificity toward Asp or Glu residues (Graf et al., 1987; Willett et al., 1995; J.J. Perona & C.S. Craik, unpubl. obs.). A better mutational strategy for specificity modification in trypsin may be the construction of libraries that instead span the distal structural elements. When coupled to strategies such as the genetic selection (Evnin et al., 1990; Perona et al., 1993a) or phage display (Corey et al., 1993; Fig. 4C) systems, it should be possible to search a large number of different structures for those providing altered specificity.

Surface loops determine subsite specificity in the trypsin-class enzymes

We have seen that the best-studied members of the chymotrypsin-like class of serine proteases each manifest primary specificity at the P1 site directly adjacent to the cleaved bond. However, there are also several enzymes in the class that possess significant specificity toward substrate residues at a greater distance in both the N- and C-terminal directions. Sequence alignments of these enzymes reveal that a number of surface loops flanking the catalytic residues are very likely to play crucial roles in determining this extended recognition selectivity (Fig. 13).

One enzyme manifesting an extended subsite specificity that is also of known tertiary structure is RMCP11 (Woodbury et al., 1978a, 1978b), a member of a homologous subclass of trypsin-like serine proteases expressed also in other granulocytes (Salvesen et al., 1987) as well as in lymphocytes (Lobe et al., 1986). RMCP11 and the related RMCP1 (which possess 73% amino acid sequence identity; LeTrong et al., 1987b) each manifest a

chymotrypsin-like primary substrate specificity but also exhibit preferences for hydrophobic amino acids in positions P2 and P3 (Yoshida et al., 1980; Powers et al., 1985). RMCP1 also has been shown to prefer hydrophobic residues at position P1' in polypeptide substrates, although the extent of the selectivity has not been established quantitatively (LeTrong et al., 1987a).

The crystal structure of uncomplexed RMCP11 has been determined at a resolution of 1.9 Å (Remington et al., 1988). This structure suggests that the enhanced substrate selectivity of the homologous RMCP1 at the P1' position is likely to be provided by the presence of a large cleft not found in the other chymotrypsin-like proteases of known structure. The cleft is formed as a consequence of an unusual conformation adopted by two surface loops that lie adjacent to the catalytic residues (Remington et al., 1988). The loops comprise residues 34–41 (loop A) and 59–64 (loop B) and are positioned such as to be capable of interacting directly with substrate residues C-terminal to the scissile bond (Fig. 13). Modeling of a substrate complex with RMCP11 suggests that loop A is most likely to directly contact the P1'-P2' substrate sites, whereas loop B plays a structural role in helping to form the cleft.

The subclass of serine proteases to which RMCP11 belongs is distinguished by the absence of the otherwise well-conserved disulfide bond linking residues 191 and 220 (LeTrong et al., 1987b). In the other enzymes, this disulfide bridges the two walls of the S1 site and likely provides a degree of structural rigidity to the cavity (Fig. 7). RMCP11 possesses a Phe residue at position 191 and a shortened loop L2 (residues 217–225) relative to chymotrypsin; each of these features is conserved within the subclass (LeTrong et al., 1987b). Modeling of a tripeptide substrate possessing Phe at position P3 shows that the aromatic ring is readily sandwiched between the side chains of Met 192 and Pro 221A and also makes van der Waals interactions with Phe 191 (Remington et al., 1988). This small hydrophobic pocket is absent in chymotrypsin owing to the presence of the Cys 191–Cys 220 disulfide bond. Thus, the crystal structure provides a plausible rationale explaining the 100-fold preference of RMCP1 and RMCP11 for Phe relative to Gly at position P3 (Yoshida et al., 1980).

A second example of extended binding site specificity is provided by the enzyme enteropeptidase (enterokinase), which functions *in vivo* to cleave the zymogen trypsinogen at position Ile 16, generating the new N-terminus required for trypsin activity (reviewed in Huber & Bode, 1978). This enzyme hydrolyzes the peptide bond directly C-terminal to the sequence (Asp)₄Lys in trypsinogen, and consequently possesses a trypsin-like specificity toward positively charged amino acids in the P1 position. The bovine and porcine enzymes exist as glycosylated disulfide-linked heterodimers comprising a heavy chain of 115 kDa and a light chain of 43 kDa (Magee et al., 1977; LaVallie et al., 1993). Chemical modification studies established that the catalytic activity and specificity of the enzyme resides in the light chain (Light & Fonseca, 1984). Most recently, cloning and expression of the light chain has revealed it to possess 35–40% sequence identity to the trypsin-like class of serine proteases (LaVallie et al., 1993). This study also demonstrated that this subunit possesses full activity toward the fluorogenic peptide substrate (Asp)₄Lys-β-naphthylamide. The presence of the heavy chain, however, endows the holoenzyme with 100-fold greater catalytic efficiency toward the cognate trypsinogen substrate (LaVallie et al., 1993).

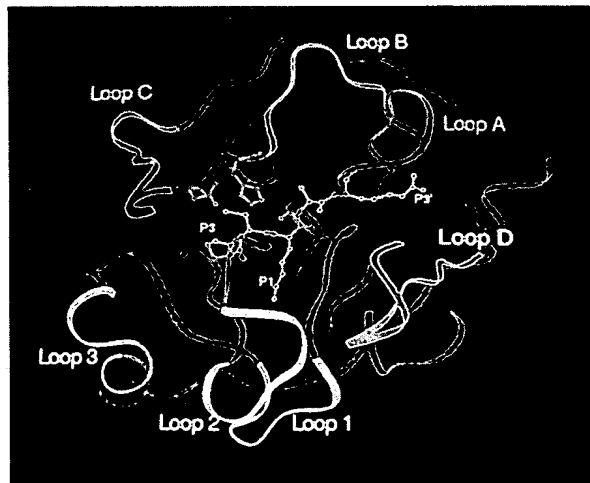


Fig. 13. Structure of trypsin, highlighting the positions of four surface loops (loops A, B, C, D) involved in determining subsite preferences among a number of the enzymes in the family. The location of these loops relative to the catalytic machinery and binding cleft may be contrasted with the position of the three loops (loops 1, 2, 3) that combine to influence specificity in the S1 site. A polypeptide substrate chain is shown in green and the catalytic triad is in yellow. It is clear that loop C is positioned to interact with substrate residues N-terminal to the scissile bond, whereas loops A and D are positioned to interact with the C-terminal amino acids on the leaving-group side of the scissile bond.

Native enteropeptidase is capable of cleaving the (Asp)₄Lys sequence in trypsinogen with a catalytic efficiency roughly 10⁴-fold greater than trypsin (Maroux et al., 1971). Mapping the sequence of the light chain of the enzyme onto the structure of trypsin indicates that the peptide Lys 96–Arg 97–Arg 98–Lys 99 (KRRK) is well positioned to play a direct role in interacting with the negatively charged aspartates occupying positions P2–P5 (LaVallie et al., 1993). This peptide comprises a portion of a surface loop located adjacent to Asp 102 (loop C; Fig. 13), on the opposing side of the catalytic triad relative to the loops A and B that form the cleft important to P1' recognition by RMCPI.

The kinetic basis for the improved specificity of enteropeptidase relative to trypsin for recognition of the (Asp)₄Lys sequence is not yet known. By analogy with the known operation of the pancreatic proteases, it would be predicted that the specificity arises at least partly from the ability of enteropeptidase to selectively accelerate the acylation rate of (Asp)₄Lys- β -naphthylamide relative to other peptidyl or to single-residue substrates. It is tempting to speculate that enteropeptidase may use a distinct structural mechanism, involving specific interactions with the aspartates, to convert substrate binding energy into a high catalytic rate. Inspection of the sequence alignment with trypsin reveals further differences at positions 215–219 at the lip of the S1 site, as well as the insertion of a residue in loop L3 (Fig. 13), each of which may be of importance to precise orientation of the (Asp)₄Lys substrate. Additionally, enteropeptidase possesses a striking 10-residue insertion between residues 58 and 59, in the surface loop B that lies directly behind the KRRK sequence of loop C (LaVallie et al., 1993; Fig. 13). Although loops B and C do not contact each other in trypsin, the much larger loop B in enteropeptidase would be capable of making interactions conceivably of importance to maintaining correct orientation of the KRRK residues.

A third example of the importance of surface loops in these enzymes relates to the inhibition of the trypsin-like tissue plasminogen activator by plasminogen activator inhibitor I (Ny et al., 1986). The interaction between TPA and PAI-1 is of importance in the regulation of the cascade of activities involved in blood clotting (Davie et al., 1991). Surface loop A of TPA (Fig. 13) possesses a high density of positively charged amino acids (residues Lys 296–His 297–Arg 298–Arg 299) that have been shown to be critical to its interaction with a negatively charged region of PAI-1 (Madison et al., 1990). This was confirmed in an elegant experiment in which loop A in the homologous enzyme thrombin was replaced with that of TPA, endowing PAI-1 susceptibility onto thrombin (Horrevoets et al., 1993). Thus, both the extended substrate specificity as well as the specificity of interaction with physiologically important inhibitors can arise from contacts with the same surface loops.

An important activity of crab collagenase is the ability to cleave native triple-helical collagen, a property not exhibited by the canonical pancreatic proteases (Eisen et al., 1973; Tsu et al., 1994). Cleavage occurs within domains of the triple-helical substrate that are relaxed from the strict Gly-Pro-Xaa repetitive sequence. Detailed examination of the cleavage sites by protein sequencing has shown that proteolysis of collagen occurs at positions that mirror the P1-site selectivity (Tsu et al., 1994). Sequence alignments of a range of serine collagenases from diverse species fails to elucidate clear amino acid similarities that might be correlated to the triple-helical specificity (Sinha et al., 1987; Sellos & Van Wormhoudt, 1992). However, the crystal structure

of collagenase complexed with the dimeric protein inhibitor ecotin has allowed construction of a model of collagen interacting with the enzyme (J.J. Perona, C.A. Tsu, R.J. Fletterick, & C.S. Craik, in prep.). Several surface loops, including loops A and D (Fig. 13), may play crucial roles in recognition of the triple helix.

Recently, a novel assay has been introduced that provides the possibility of assaying relative preferences at positions on the leaving-group side of the scissile bond (Schellenberger et al., 1993). In an initial study, the S1' subsite specificities of trypsin and chymotrypsin from cow and rat were determined by monitoring the reverse reaction of peptide hydrolysis. Acyl transfer was measured to a mixture of 21 peptide nucleophiles of the general structure H-Xaa-Ala-Ala-Ala-NH₂; the decrease in concentration of each nucleophile was monitored by HPLC and represents a measure of the ability of that substrate to compete with water for attack on the acyl enzyme. Chymotrypsin hydrolyzes substrates possessing Arg and Lys at the substrate P1' position roughly 10-fold more rapidly than does trypsin; this selectivity is attributable to the presence of additional negatively charged residues in two adjacent surface loops (see below). Trypsin exhibits a slight preference for hydrophobic amino acids at this position, relative to chymotrypsin. The data confirm the relative lack of specificity of each enzyme at this position. Application of the methodology to crab collagenase showed a 30-fold preference for P1'-Arg residues; an Arg is also found on the C-terminal side of several of the collagen cleavage sites of the enzyme (Tsu et al., 1994). Data have also been obtained for specificities at the subsites S1'–S3' for trypsin, chymotrypsin, α -lytic protease, and the cercarial protease from *Schistosoma mansoni*; in these cases, relative cleavage rates varied by factors of up to 10²-fold (Schellenberger et al., 1994).

It is clear from the many known structures of chymotrypsin-like serine proteases that loop C is invariably positioned to directly contact the extended substrate on the N-terminal side, whereas loops A and D interact on the leaving group side. By contrast, loop B appears less likely to be involved in direct contacts but instead is positioned to stabilize the primary interactions made by the more forward loops (Fig. 13). Depending on the size and conformation of this loop in different enzymes, it might in principle be able to stabilize either loop A or C. A final example of specificity modification in this class involves loop D: introduction of histidine residues at the N- and C-terminal ends of this loop confers metal-dependent specificity for histidine at the P2' substrate position onto rat trypsin (Willett et al., 1995). In general, because subsite specificity of chymotrypsin-like proteases is modulated by surface loops rather than by core secondary structure elements, the prospects for engineering novel specificities, and for the development of "restriction proteases" that might recognize substrate sites from P5 to P2', seem hopeful.

Conclusions and future directions

One of the questions addressed in these studies is the role of water molecules in mediating enzyme–ligand interactions. Crystal structures of wild-type and variant enzymes complexed with substrate analogs, together with the measurement of affinity constants, allows deduction of the importance of particular interactions. In the recognition of basic Lys and Arg substrate side chains by Asp 189 of trypsin, the conclusion is that a water-mediated interaction can provide a comparable free energy gain to a direct contact (Perona et al., 1993c). These studies have im-

plications to understanding protein-nucleic acid interactions. For example, the crystal structures of the *trp* repressor-operator complex, and of the uncomplexed operator DNA, suggest a crucial role for water-mediated interactions in providing DNA sequence specificity because no direct contacts with base functional groups are observed (Otwinowski et al., 1988; Shakked et al., 1994). Although a second-site reversion analysis of the operator DNA further implied a key role for the intervening waters, it was clear that a structural analysis of the modified complexes is still required (Joachimiak et al., 1994). Such an analysis of the charge-charge interactions in the trypsin S1 site shows more definitively that a specificity-determining role for solvent is in principle possible. A similar study of the *trp* repressor and of other systems is warranted, to address the extent to which this phenomenon may be dependent on the local structural context.

Another fundamental question concerns the design of enzyme structures to provide different degrees of flexibility to the substrate binding site. The comparison of trypsin and α -lytic protease offers an excellent opportunity to address this issue. Thus far, it appears from both kinetic and structural analysis of mutants that the trypsin pocket may be considerably more rigid. However, the two structures are homologous so that the degree of difference in the surrounding scaffolds is relatively small. Thus, the problem may be manageable: which specific interactions bridging the primary and secondary shell residues are most critical for determining flexibility? Are residues located even more distant also important? An excellent test of our understanding would be the construction of a trypsin variant with chymotryptic specificity, which possessed far fewer than the 15 alterations of Tr \rightarrow Ch[S1+L1+L2+Y172W]. If indeed the conformation of Gly 216 is crucial to P1-site specificity, then the problem reduces to adding certain key mutations to D189S such that Gly 216 can reorient upon substrate binding, as it is observed to do in α -lytic protease (Bone et al., 1991; Fig. 9). A deeper understanding of flexibility would have clear application to protein folding and stability as well (Rose & Creamer, 1994).

The degree to which a substrate binding cleft is inherently deformable may be an important parameter governing the ease with which specificity modification can be effected. Prior to the advent of site-directed mutagenesis, it appeared possible that even conservative amino acid changes might cause highly deleterious long-range structural effects. We now know that most substitutions are absorbed locally and that the majority of protein structural contexts therefore have some ability to deform. Protein folding and stability often are not greatly perturbed even by very challenging mutations. The sensitivity of enzyme activity to precise substrate positioning might alternatively suggest that mutation of the binding site would usually result in low catalytic activity. However, this appears not to be the case: among the well-studied binding pockets considered here, the subtilisin S1 and S4 sites, as well as the α -lytic protease S1 site, each are readily modified to alter specificity with only limited local substitutions. Only the trypsin S1 site requires extensive nonlocal changes.

Another reason for the difficulty in modifying trypsin substrate specificity could be that the charge-charge interactions present in a trypsin transition-state complex require a precise electrostatic environment not readily altered (Hwang & Warshel, 1988). The electrostatic potential is presently the least understood force shaping enzyme structure and activity; it is also the only one that operates over large distances. Considerable efforts

are underway to improve empirical forcefields, so that catalytic free energies can be accurately estimated directly from structural models. Serine proteases are a favored system in these studies owing to the large database of structure-activity information (Bash et al., 1987; Rao et al., 1987; Caldwell et al., 1991; Mizushima et al., 1991; Wilson et al., 1991). Further mutational analysis will thus also be invaluable in providing a testbed for new algorithms. Greater insight into the connection between structure and energetics will lead to much better predictive ability regarding the consequences of mutation. This improved insight, together with the innovative technologies for the generation and screening of large libraries, may soon result in the creation of new, highly efficient proteases possessing a broad range of useful properties.

Acknowledgments

We thank Prof. R. Fletterick for helpful discussions. This work was supported by NSF grants MCB-9219806 and BCS-9119237 to C.S.C. and NIH postdoctoral NRSA award GM 13818-03 to J.J.P.

References

- Abrahmsen L, Tom J, Burnier J, Butcher KA, Kossiakoff A, Wells JA. 1991. Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution. *Biochemistry* 30:4151-4159.
- Bachovchin WW, Roberts JD. 1978. Nitrogen-15 nuclear magnetic resonance spectroscopy. The state of histidine in the catalytic triad of α -lytic protease. Implications for the charge-relay mechanism of peptide-bond cleavage by serine proteases. *J Am Chem Soc* 100:8041-8047.
- Bash PA, Singh UC, Langridge R, Kollman PA. 1987. Free energy calculations by computer simulation. *Science* 236:564-566.
- Bauer CA. 1978. Active centers of *Streptomyces griseus* protease 1, *Streptomyces griseus* protease 3, and α -chymotrypsin: Enzyme-substrate interactions. *Biochemistry* 17:375-380.
- Bauer CA, Brayer GD, Sielecki AR, James MNG. 1981. Active site of α -lytic protease. Enzyme-substrate interactions. *Eur J Biochem* 120:289-294.
- Bauer CA, Thompson RC, Blout ER. 1976. The active centers of *Streptomyces griseus* protease 3 and α -chymotrypsin: Enzyme-substrate interactions remote from the scissile bond. *Biochemistry* 15:1291-1295.
- Bazan JF, Fletterick RJ. 1990. Structural and catalytic models of trypsin-like viral proteases. *Semin Virol* 1:311-322.
- Bech LM, Sørensen SB, Breddam K. 1992. Mutational replacements in subtilisin 309. Val 104 has a modulating effect on the P4 substrate preference. *Eur J Biochem* 209:869-874.
- Bech LM, Sørensen SB, Breddam K. 1993. Significance of hydrophobic S4-P4 interactions in subtilisin 309 from *Bacillus lentus*. *Biochemistry* 32:2845-2852.
- Bender ML, Killheffer JV. 1973. Chymotrypsins. *CRC Crit Rev Biochem* 1:149-199.
- Betzel C, Klupsch S, Papendorf G, Hastrup S, Branner S, Wilson KS. 1992. Crystal structure of the alkaline proteinase savinase from *Bacillus lentus* at 1.4 Å resolution. *J Mol Biol* 223:427-445.
- Betzel C, Pal GP, Saenger W. 1988. Three-dimensional structure of proteinase K at 0.15-nm resolution. *Eur J Biochem* 178:155-171.
- Betzel C, Singh TP, Visanji M, Fittkau S, Saenger W, Wilson KS. 1993. Structure of the complex of proteinase K with a substrate analogue hexapeptide inhibitor at 2.2-Å resolution. *J Biol Chem* 268:15854-15858.
- Blow DM. 1976. Structure and mechanism of chymotrypsin. *Acc Chem Res* 9:145-152.
- Blow DM, Birktoft JJ, Hartley BS. 1969. Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature* 221:337-340.
- Bode W, Chen Z, Bartels K, Kutzbach C, Schmidt-Kastner G, Bartunik H. 1983. Refined 2 Å X-ray crystal structure of porcine pancreatic kallikrein A, a specific trypsin-like serine proteinase—Crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine trypsin. *J Mol Biol* 164:237-282.
- Bode W, Mayr I, Baumann U, Huber R, Stone SR, Hofsteenge J. 1989a. The refined 1.9 Å crystal structure of human α -thrombin: Interaction

- with D-Phe-Pro-Arg chloromethyl ketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J* 8:3467-3475.
- Bode W, Meyer E Jr, Powers JC. 1989b. Human leukocyte and porcine pancreatic elastase: X-ray crystal structures, mechanism, substrate specificity, and mechanism-based inhibitors. *Biochemistry* 28:1951-1963.
- Bode W, Papamokos E, Musil D, Seemuller U, Fritz H. 1986a. Refined 1.2 Å crystal structure of the complex formed between subtilisin Carlsberg and the inhibitor eglin C. Molecular structure of eglin and its detailed interaction with subtilisin. *EMBO J* 5:813-818.
- Bode W, Walter J, Huber R, Wenzel HR, Tschesche H. 1984. The refined 2.2 Å X-ray crystal structure of the ternary complex formed by bovine trypsinogen, valine-valine and the Arg 15 analogue of bovine pancreatic trypsin inhibitor. *Eur J Biochem* 144:185-190.
- Bode W, Wei AZ, Huber R, Meyer EF Jr, Travis J, Neumann S. 1986b. X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO J* 5:2453-2458.
- Bone R, Agard DA. 1991. Mutational remodeling of enzyme specificity. *Methods Enzymol* 202:643-671.
- Bone R, Frank D, Kettner CA, Agard DA. 1989a. Structural analysis of specificity: α -lytic protease complexes with analogues of reaction intermediates. *Biochemistry* 28:7600-7609.
- Bone R, Fujishige A, Kettner CA, Agard DA. 1991. Structural basis for broad specificity in α -lytic protease mutants. *Biochemistry* 30:10388-10398.
- Bone R, Shenvi AB, Kettner CA, Agard DA. 1987. Serine protease mechanism: Structure of an inhibitory complex of α -lytic protease and a tightly bound peptide boronic acid. *Biochemistry* 26:7609-7614.
- Bone R, Silen JL, Agard DA. 1989b. Structural plasticity broadens the specificity of an engineered protease. *Nature* 339:191-195.
- Braxton S, Wells JA. 1991. The importance of a distal hydrogen bonding group in stabilizing the transition state in subtilisin BPN'. *J Biol Chem* 266:11797-11800.
- Brayer GD, Delbaere LTJ, James MNG. 1979. Molecular structure of the α -lytic protease from *Myxobacter 495* at 2.8 Å resolution. *J Mol Biol* 131:743-775.
- Brenner C, Fuller RS. 1992. Structural and enzymatic characterization of a purified, prohormone-processing enzyme: Secreted, soluble Kex2 protease. *Proc Natl Acad Sci USA* 89:922-926.
- Bryan P, Pantoliano MW, Quill SG, Hsiao HY, Poulos T. 1986. Site-directed mutagenesis and the role of the oxyanion hole in subtilisin. *Proc Natl Acad Sci USA* 83:3743-3745.
- Caldwell JW, Agard DA, Kollman PA. 1991. Free energy calculations on binding and catalysis by α -lytic protease: The role of substrate size in the P1 pocket. *Proteins Struct Funct Genet* 10:140-148.
- Caputo A, James MNG, Powers JC, Hudig D, Bleackley RC. 1994. Conversion of the substrate specificity of mouse proteinase granzyme B. *Struct Biol* 1:364-367.
- Carter P, Abrahmsen L, Wells JA. 1991. Probing the mechanism and improving the rate of substrate-assisted catalysis in subtilisin BPN'. *Biochemistry* 30:6142-6148.
- Carter P, Nilsson B, Burnier JP, Burdick D, Wells JA. 1989. Engineering subtilisin BPN' for site-specific proteolysis. *Proteins Struct Funct Genet* 6:240-248.
- Carter P, Wells JA. 1987. Engineering enzyme specificity by "substrate-assisted catalysis." *Science* 237:394-399.
- Carter P, Wells JA. 1988. Dissecting the catalytic triad of a serine protease. *Nature* 332:564-568.
- Carter P, Wells JA. 1990. Functional interactions among catalytic residues in subtilisin BPN'. *Proteins Struct Funct Genet* 7:335-342.
- Chasan R, Anderson KV. 1989. The role of Easter, an apparent serine protease, in organizing the dorsal-ventral pattern of the *Drosophila* embryo. *Cell* 56:391-400.
- Corey DR, Craik CS. 1992. An investigation into the minimum requirements for peptide hydrolysis by mutation of the catalytic triad of trypsin. *J Am Chem Soc* 114:1784-1790.
- Corey DR, McGrath ME, Vasquez JR, Fletterick RJ, Craik CS. 1992. An alternate geometry for the catalytic triad of serine proteases. *J Am Chem Soc* 114:4906-4907.
- Corey DR, Shiau AK, Yang Q, Janowski B, Craik CS. 1993. Trypsin display on the surface of bacteriophage. *Gene* 128:129-134.
- Craik CS, Largman C, Fletcher T, Roczniak S, Barr PJ, Fletterick RJ, Rutter WJ. 1985. Redesigning trypsin: Alteration of substrate specificity. *Science* 228:291-297.
- Craik CS, Roczniak S, Largman C, Rutter WJ. 1987. The catalytic role of the active site aspartic acid in serine proteases. *Science* 237:909-913.
- Creemers JWM, Siezen RJ, Roebroek AJM, Ayoubi TAY, Huylebroeck D, Van de Ven WJM. 1993. Modulation of furin-mediated proprotein processing activity by site-directed mutagenesis. *J Biol Chem* 268:21826-21834.
- Dancer SJ, Garratt R, Saldanha J, Jhoti H, Evans R. 1990. The epidermolytic toxins are serine proteases. *FEBS Lett* 268:129-132.
- Davie EW, Fujikawa K, Kisiel W. 1991. The coagulation cascade: Initiation, maintenance and regulation. *Biochemistry* 30:10363-10370.
- Delbaere LTJ, Brayer GD, James MNG. 1979. The 2.8 Å resolution structure of *Streptomyces griseus* protease B and its homology with α -chymotrypsin and *Streptomyces griseus* protease A. *Can J Biochem* 57:135-144.
- Dixon GH, Go S, Neurath H. 1956. Peptides combined with ^{14}C -diisopropyl phosphoryl following degradation of ^{14}C -DIP-trypsin with α -chymotrypsin. *Biochim Biophys Acta* 19:193-200.
- Drapeau GR. 1978. The primary structure of staphylococcal protease. *Can J Biochem* 56:534-544.
- Eder J, Rheinhecker M, Fersht AR. 1993. Hydrolysis of small peptide substrates parallels binding of chymotrypsin inhibitor 2 for mutants of subtilisin BPN'. *FEBS Lett* 335:349-352.
- Eisen AZ, Henderson KO, Jeffrey JJ, Bradshaw RA. 1973. A collagenolytic protease from the hepatopancreas of the fiddler crab, *Uca pugilator*. Purification and properties. *Biochemistry* 12:1814-1822.
- Epstein DM, Abells RH. 1992. Role of serine 214 and tyrosine 171, components of the S2 subsite of α -lytic protease, in catalysis. *Biochemistry* 31:11216-11223.
- Estell DA, Graycar TP, Miller JV, Powers DB, Burnier JP, Ng PG, Wells JA. 1986. Probing steric and hydrophobic effects on enzyme-substrate interactions by protein engineering. *Science* 233:659-663.
- Evin LB, Vasquez JR, Craik CS. 1990. Substrate specificity of trypsin investigated by using a genetic selection. *Proc Natl Acad Sci USA* 87:6659-6663.
- Fujinaga M, Delbaere LTJ, Brayer GD, James MNG. 1985. Refined structure of α -lytic protease at 1.7 Å resolution. Analysis of hydrogen bonding and solvent structure. *J Mol Biol* 183:479-502.
- Fujinaga M, James MNG. 1987. Rat submaxillary gland serine protease, tonin—Structure solution and refinement at 1.8 Å resolution. *J Mol Biol* 195:373-396.
- Fuller RS, Brake A, Thorner J. 1989. Yeast prohormone processing enzyme (KEX2 gene product) is a Ca^{2+} -dependent serine protease. *Proc Natl Acad Sci USA* 86:1434-1438.
- Graf L, Craik CS, Pathy A, Roczniak S, Fletterick RJ, Rutter WJ. 1987. Selective alteration of substrate specificity by replacement of aspartic acid-189 with lysine in the binding pocket of trypsin. *Biochemistry* 26:2616-2623.
- Graf L, Jancso A, Szilagyi L, Hegyi G, Pinter K, Naray-Szabo G, Hepp J, Medzihradsky K, Rutter WJ. 1988. Electrostatic complementarity in the substrate binding pocket of trypsin. *Proc Natl Acad Sci USA* 85:4961-4965.
- Graham LD, Haggett KD, Jennings PA, LeBrocq DS, Whittaker RG, Schober PA. 1993. Random mutagenesis of the substrate binding site of a serine protease can generate enzymes with increased activities and altered primary specificities. *Biochemistry* 32:6250-6258.
- Grant GA, Eisen AZ. 1980. Substrate specificity of the collagenolytic serine protease from *Uca pugilator*: Studies with noncollagenous substrates. *Biochemistry* 19:6089-6095.
- Grant GA, Henderson KO, Eisen AZ, Bradshaw RA. 1980. Amino acid sequence of a collagenolytic protease from the hepatopancreas of the fiddler crab, *Uca pugilator*. *Biochemistry* 19:4653-4659.
- Greer J. 1990. Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins Struct Funct Genet* 7:317-334.
- Grøn H, Breddam K. 1992. Interdependency of the binding sites in subtilisin. *Biochemistry* 31:8967-8971.
- Grøn H, Meldal M, Breddam K. 1992. Extensive comparison of the substrate preferences of two subtilisins as determined with peptide substrates which are based on the principle of intramolecular quenching. *Biochemistry* 31:6011-6018.
- Gros P, Betzel C, Dauter Z, Wilson KS, Hol WGJ. 1989. Molecular dynamics refinement of a thermolysin-eglin-c complex at 1.98 Å resolution and comparison of two crystal forms that differ in calcium content. *J Mol Biol* 210:347-367.
- Guo J, Huang W, Scanlan TS. 1994. Kinetic and mechanistic characterization of a highly active hydrolytic antibody: Evidence for the formation of an acyl intermediate. *J Am Chem Soc* 116:6062-6069.
- Harper JW, Cook RR, Roberts CJ, McLaughlin BJ, Powers JC. 1984. Active site mapping of the serine proteases human leukocyte elastase, cathepsin G, porcine pancreatic elastase, rat mast cell proteases I and II, bovine chymotrypsin A α and *S. aureus* protease V-8 using tripeptide thio-benzyl ester substrates. *Biochemistry* 23:2995-3002.

- Hedstrom L, Farr-Jones S, Kettner CA, Rutter WJ. 1994a. Converting trypsin to chymotrypsin: Ground state binding does not determine substrate specificity. *Biochemistry* 33:8764-8769.
- Hedstrom L, Perona JJ, Rutter WJ. 1994b. Converting trypsin to chymotrypsin: Residue 172 is a substrate specificity determinant. *Biochemistry* 33:8757-8763.
- Hedstrom L, Szilagyi L, Rutter WJ. 1992. Converting trypsin to chymotrypsin: The role of surface loops. *Science* 255:1249-1253.
- Higaki J, Evnin LB, Craik CS. 1989. Introduction of a cysteine protease active site into trypsin. *Biochemistry* 28:9256-9263.
- Higaki JN, Haymore BL, Chen S, Fletterick R, Craik CS. 1990. Regulation of serine protease activity by an engineered metal switch. *Biochemistry* 29:8582-8586.
- Horrevoets AJG, Tans G, Smilde AE, van Zonneveld AJ, Pannekoek H. 1993. Thrombin-variable region 1. Evidence for the dominant contribution of VR1 of serine proteases to their interaction with plasminogen activator inhibitor 1. *J Biol Chem* 268:779-782.
- Huber R, Bode W. 1978. Structural basis of the activation and action of trypsin. *Acc Chem Res* 11:114-122.
- Hwang JK, Warshel A. 1988. Why ion pair reversal by protein engineering is unlikely to succeed. *Nature* 334:270-272.
- Jackson DY, Burnier J, Quan C, Stanley M, Tom J, Wells JA. 1994. A designed peptide ligase for total synthesis of ribonuclease A with unnatural catalytic residues. *Science* 266:243-247.
- Jackson SE, Fersht AR. 1993. Contribution of long-range electrostatic interactions to the stabilization of the catalytic transition state of the serine protease subtilisin BPN'. *Biochemistry* 32:13909-13916.
- James MNG. 1976. Relationship between the structures and activities of some microbial serine proteases. II. Comparison of the tertiary structures of microbial and pancreatic serine proteases. In: Ribbons DW, Brew K, eds. *Proteolysis and physiological regulation*. New York: Academic Press. pp 125-142.
- Joachimiak A, Haran TE, Sigler PB. 1994. Mutagenesis supports water mediated recognition in the *trp* repressor-operator system. *EMBO J* 13:367-372.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.
- Kahne D, Still WC. 1988. Hydrolysis of a peptide bond in neutral water. *J Am Chem Soc* 110:7529-7534.
- Kettner CA, Shenoi AB. 1984. Inhibition of the serine proteases leukocyte elastase, pancreatic elastase, cathepsin G and chymotrypsin by peptide boronic acids. *J Biol Chem* 259:15106-15114.
- Kossiakoff AA, Spencer SA. 1981. Direct determination of the protonation states of aspartic acid-102 and histidine-57 in the tetrahedral intermediate of the serine proteases: Neutron structure of trypsin. *Biochemistry* 20:6462-6473.
- Kraut J. 1977. Serine proteases: Structure and mechanism of catalysis. *Annu Rev Biochem* 46:331-358.
- LaVallie ER, Rehemtulla A, Racie LA, DiBlasio EA, Ferentz C, Grant KL, Light A, McCoy JM. 1993. Cloning and functional expression of a cDNA encoding the catalytic subunit of bovine enterokinase. *J Biol Chem* 268:23311-23317.
- Lazure C, Scidah NG, Pelaprat D, Chretien M. 1983. Proteases and post-translational processing of prohormones: A review. *Can J Biochem Cell Biol* 61:501-515.
- LeTrong H, Neurath H, Woodbury RG. 1987a. Substrate specificity of the chymotrypsin-like protease in secretory granules isolated from rat mast cells. *Proc Natl Acad Sci USA* 84:364-367.
- LeTrong H, Parmelee DD, Walsh KA, Neurath H, Woodbury RG. 1987b. Amino acid sequence of rat mast cell protease I (chymase). *Biochemistry* 26:6988-6994.
- Liao D, Breddam K, Sweet RM, Bullock T, Remington SJ. 1992. Refined atomic model of wheat serine carboxypeptidase II at 2.2 Å resolution. *Biochemistry* 31:9796-9812.
- Liao D, Remington SJ. 1990. Structure of wheat serine carboxypeptidase II at 3.5 Å resolution. A new class of serine protease. *J Biol Chem* 265:6528-6531.
- Light A, Fonseca P. 1984. The preparation and properties of the catalytic subunit of bovine enterokinase. *J Biol Chem* 259:13195-13198.
- Lobe CG, Finlay BB, Paranchych W, Paetkau VH, Bleackley RC. 1986. Novel serine proteases encoded by two cytotoxic T lymphocyte-specific genes. *Science* 232:858-861.
- Loewenthal R, Sancho J, Reinikainen T, Fersht AR. 1993. Long-range surface charge-charge interactions in proteins. Comparison of experimental results with calculations from a theoretical method. *J Mol Biol* 232:574-583.
- Madison EL, Goldsmith EJ, Gerard RD, Gething MJH, Sambrook JF, Bassel-Duby RS. 1990. Amino acid residues that affect interaction of tissue-type plasminogen activator with plasminogen activator inhibitor 1. *Proc Natl Acad Sci USA* 87:3530-3533.
- Magee AI, Grant DAW, Hermon-Taylor J. 1977. The apparent molecular weights of human intestinal aminopeptidase, enterokinase and maltase in native duodenal fluid. *Biochem J* 165:583-585.
- Markland FS, Smith EL. 1971. Subtilisins: Primary structure, chemical and physical properties. In: Boyer PD, ed. *The enzymes*, vol 3. New York: Academic Press. pp 516-608.
- Markley JL. 1979. Catalytic groups of serine proteases. NMR investigations. In: Shulman RG, ed. *Biological applications of magnetic resonance*. New York: Academic Press. pp 397-461.
- Maroux S, Barratti J, Desnuelle P. 1971. Purification and specificity of porcine enterokinase. *J Biol Chem* 246:5031-5039.
- Matthews BW. 1977. X-ray structure of proteins. In: Neurath H, Hill RL, eds. *The proteins*, vol 3. New York: Academic Press. pp 404-590.
- Matthews BW, Sigler PB, Henderson R, Blow DM. 1967. Three-dimensional structure of tosyl-α-chymotrypsin. *Nature* 214:652-656.
- Matthews DJ, Wells JA. 1993. Substrate phase: Selection of protease substrates by monovalent phage display. *Science* 260:1113-1117.
- Matthews G, Shennan KI, Seal AJ, Taylor NA, Colman A, Docherty K. 1994. Autocatalytic maturation of the proenzyme convertase PC2. *J Biol Chem* 269:588-592.
- McGrath ME, Haymore BL, Summers NL, Craik CS, Fletterick RJ. 1993. Structure of an engineered, metal-actuated switch in trypsin. *Biochemistry* 32:1914-1919.
- McGrath ME, Vasquez JR, Craik CS, Yang AS, Honig B, Fletterick RJ. 1992. Perturbing the polar environment of Asp 102 in trypsin: Consequences of replacing conserved Ser 214. *Biochemistry* 31:3059-3064.
- McGrath M, Wilke ME, Higaki JN, Craik CS, Fletterick R. 1989. Crystal structures of two engineered thiol trypsin. *Biochemistry* 28:9264-9270.
- McPhalen CA, James MNG. 1988. Structural comparison of two serine proteinase-protein inhibitor complexes: Eglin C-subtilisin Carlsberg and Cl-2-subtilisin Novo. *Biochemistry* 27:6582-6598.
- Mizushima N, Spellmeyer D, Hirono S, Pearlman D, Kollman P. 1991. Free energy perturbation calculations on binding and catalysis after mutating threonine 220 in subtilisin. *J Biol Chem* 266:11801-11809.
- Mortenson UH, Remington SJ, Breddam K. 1994. Site-directed mutagenesis on (serine) carboxypeptidase Y: A hydrogen-bond network stabilizes the transition state by interaction with the C-terminal carboxylate of the substrate. *Biochemistry* 33:508-517.
- Moult J, Sussman F, James MNG. 1985. Electron density calculations as an extension of protein structure refinement. *Streptomyces griseus* protease A at 1.5 Å resolution. *J Mol Biol* 182:555-566.
- Murphy MEP, Moult J, Bleackley RC, Gershenfeld H, Weissman IL, James MNG. 1988. Comparative molecular model building of two serine proteinases from cytotoxic T lymphocytes. *Protein Struct Funct Genet* 4:190-204.
- Nakatsuka T, Sasaki T, Kaiser ET. 1987. Peptide segment coupling catalyzed by the semisynthetic enzyme thiolsubtilisin. *J Am Chem Soc* 109:3808-3810.
- Narajana SV, Carson M, el-Kabbiani O, Kilpatrick JM, Moore D, Chen X, Bugg CE, DeLucas LJ. 1994. Structure of human factor D. A complement system protein at 2.0 Å resolution. *J Mol Biol* 235:695-708.
- Navia MA, McKeever BM, Springer JP, Lin TY, Williams HR, Fluder EM, Dorn CP, Hoogsteen K. 1989. Structure of human neutrophil elastase in complex with a peptide chloromethyl ketone inhibitor at 1.84 Å resolution. *Proc Natl Acad Sci USA* 86:7-11.
- Neurath H. 1984. Evolution of proteolytic enzymes. *Science* 224:350-357.
- Neurath H. 1985. Proteolytic enzymes, past and present. *Fed Proc* 44:2907-2913.
- Nienaber VL, Breddam K, Birktoft JJ. 1993. A glutamic acid specific serine protease utilizes a novel histidine triad in substrate binding. *Biochemistry* 32:11469-11475.
- Ny T, Sawdey M, Lawrence D, Millan JL, Lorsukoff DJ. 1986. Cloning and sequence of a cDNA coding for the human beta-migrating endothelial-cell-type plasminogen activator inhibitor. *Proc Natl Acad Sci USA* 83:6776-6780.
- Odake S, Kam CM, Narasimhan L, Poe M, Blake JT, Krahenbuhl O, Tschopp J, Powers JC. 1991. Human and murine cytotoxic T-lymphocyte serine proteases: Subsite mapping with peptide thioester substrates and inhibition of enzyme activity and cytotoxicity by isocoumarins. *Biochemistry* 30:2217-2227.
- Ollis DL, Cheah E, Cygler M, Dykstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A. 1992. The alpha/beta hydrolase fold. *Protein Eng* 5:197-211.
- Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, Mar-

- morstein RQ, Luisi BF, Sigler PB. 1988. Crystal structure of *trp* repressor/operator at atomic resolution. *Nature* 335:321-327.
- Padmanabhan K, Padmanabhan KP, Tulinsky A, Park CH, Bode W, Huber R, Blankenship DT, Cardin AD, Kiesel W. 1993. Structure of human ds(1-45) factor Xa at 2.2 Å resolution. *J Mol Biol* 232:947-966.
- Perona JJ, Evnin LB, Craik CS. 1993a. A genetic selection elucidates structural determinants of arginine versus lysine specificity in trypsin. *Gene* 137:121-126.
- Perona JJ, Hedstrom L, Rutter WJ, Fletterick RJ. 1995. Structural origins of substrate discrimination in trypsin and chymotrypsin. *Biochemistry* 34:1489-1499.
- Perona JJ, Hedstrom L, Wagner R, Rutter WJ, Craik CS, Fletterick RJ. 1994. Exogenous acetate reconstitutes the enzymatic activity of Asp 189 Ser trypsin. *Biochemistry* 33:3252-3259.
- Perona JJ, Tsu CA, Craik CS, Fletterick RJ. 1993b. Crystal structures of rat anionic trypsin complexed with the protein inhibitors APPI and BPTI. *J Mol Biol* 230:919-933.
- Perona JJ, Tsu CA, McGrath ME, Craik CS, Fletterick RJ. 1993c. Relocating a negative charge in the binding pocket of trypsin. *J Mol Biol* 230:934-949.
- Polgar L. 1989. Structure and function of serine proteases. In: *Mechanisms of protease action*. Boca Raton, Florida: CRC Press. Chapter 3.
- Polgar L. 1991. pH-dependent mechanism in the catalysis of prolyl endopeptidase from pig muscle. *Eur J Biochem* 197:441-447.
- Poulos TL, Alden RA, Freer ST, Birktoft JJ, Kraut J. 1976. Polypeptide halo-methyl ketones bind to serine proteases as analogs of the tetrahedral intermediate. *J Biol Chem* 251:1097-1103.
- Powers JC, Tanaka T, Harper JW, Minematsu Y, Barker L, Lincoln D, Crumley KV, Fraki JE, Schechter NM, Lazarus GG, Nakajima K, Nakashino K, Neurath H, Woodbury RG. 1985. Mammalian chymotrypsin-like enzymes. Comparative reactivities of rat mast cell proteases, human and dog skin chymases, and human cathepsin G with peptide 4-nitroanilide substrates and with peptide chloromethyl ketone and sulfonyl fluoride inhibitors. *Biochemistry* 24:2048-2058.
- Rao SN, Singh UC, Bash PA, Kollman PA. 1987. Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature* 328:551-554.
- Read RJ, James MNG. 1988. Refined crystal structure of *Streptomyces griseus* trypsin at 1.7 Å resolution. *J Mol Biol* 200:523-551.
- Rechmulla A, Barr PJ, Rhodes CJ, Kaufman RJ. 1993. PACE4 is a member of the mammalian propeptidase family that has overlapping but not identical substrate specificity to PACE. *Biochemistry* 32:11586-11590.
- Remington SJ, Woodbury RG, Reynolds RA, Matthews BW, Neurath H. 1988. The structure of rat mast cell protease at 1.9-Å resolution. *Biochemistry* 27:8097-8105.
- Rheinhecker M, Baker G, Eder J, Fersht AR. 1993. Engineering a novel specificity in subtilisin BPN'. *Biochemistry* 32:1199-1203.
- Rheinhecker M, Eder J, Pandey PS, Fersht AR. 1994. Variants of subtilisin BPN' with altered specificity profiles. *Biochemistry* 33:221-225.
- Robertus JD, Alden RA, Birktoft JJ, Kraut J, Powers JC, Wilcox PE. 1972a. An X-ray crystallographic study of the binding of peptide chloromethyl ketone inhibitors to subtilisin BPN'. *Biochemistry* 11:2439-2449.
- Robertus JD, Kraut J, Alden RA, Birktoft J. 1972b. Subtilisin: A stereochemical mechanism involving transition-state stabilization. *Biochemistry* 11:4293-4303.
- Rose GD, Creamer TP. 1994. Protein folding: Predicting predicting. *Proteins Struct Funct Genet* 19:1-3.
- Ruhlmann A, Kukla D, Schwager P, Bartels K, Huber R. 1973. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. *J Mol Biol* 77:417-436.
- Russell AJ, Thomas PG, Fersht AR. 1987. Electrostatic effects on modification of charged groups in the active site cleft of subtilisin by protein engineering. *J Mol Biol* 193:803-813.
- Salvesen G, Farley D, Shuman J, Przybyla A, Reilly C, Travis J. 1987. Molecular cloning of human cathepsin G: Structural similarity to mast cell and cytotoxic T lymphocyte proteinases. *Biochemistry* 26:2289-2293.
- Schechter I, Berger A. 1968. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* 27:157-162.
- Schellenberger V, Turck CW, Hedstrom L, Rutter WJ. 1993. Mapping the S' subsites of serine proteases using acyl transfer to mixtures of peptide nucleophiles. *Biochemistry* 32:4349-4353.
- Schellenberger V, Turck CW, Rutter WJ. 1994. Role of the S' subsites in serine protease catalysis. Active-site mapping of rat chymotrypsin, rat trypsin, α -lytic protease and cercarial protease from *Schistosoma mansoni*. *Biochemistry* 33:4251-4257.
- Seidah NG, Day R, Marcinkiewicz M, Benjannet S, Chretien M. 1991. Mammalian neural and endocrine pro-protein and pro-hormone convertases belonging to the subtilisin family of serine proteases. *Enzyme* 45:271-284.
- Sellos D, Van Wormhoudt A. 1992. Molecular cloning of a cDNA that encodes a serine protease with chymotryptic and collagenolytic activities in the hepatopancreas of the shrimp *Penaeus vanamei* (Crustacea, Decapoda). *FEBS Lett* 309:219-224.
- Shakked Z, Guzikevich-Guerstein G, Frolow F, Rabinovich D, Joachimiak A, Sigler PB. 1994. Determinants of repressor-operator recognition from the structure of the *trp* operator binding site. *Nature* 368:469-473.
- Shaw E, Mares-Guia M, Cohen W. 1965. Evidence for an active site histidine in trypsin through use of a specific reagent, 1-chloro-3-tosylamido-7-amino-2-heptanone, the chloromethyl ketone derived from N^α-tosyl-L-lysine. *Biochemistry* 4:2219-2226.
- Siezen RJ, Bruinenberg PG, Vos P, van Alen-Boerrigter I, Nijhuis M, Altling AC, Exterkate FA, de Vos WM. 1993. Engineering of the substrate-binding region of the subtilisin-like, cell-envelope proteinase of *Lactococcus lactis*. *Protein Eng* 6:927-937.
- Siezen RJ, de Vos WM, Leunissen JAM, Dijkstra BW. 1991. Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteases. *Protein Eng* 4:717-719.
- Sinha S, Watorek W, Karr S, Giles J, Bode W, Travis J. 1987. Primary structure of human neutrophil elastase. *Proc Natl Acad Sci USA* 84:2228-2232.
- Smeekens SP, Avruch AS, LaMendola J, Chan SJ, Steiner DF. 1991. Identification of a cDNA encoding a second putative prohormone convertase related to PC2 in AtT20 cells and islets of Langerhans. *Proc Natl Acad Sci USA* 88:340-344.
- Smith CL, DeLotto R. 1994. Ventralizing signal determined by protease activation in *Drosophila* embryogenesis. *Nature* 368:548-551.
- Sørensen SB, Bech LM, Meldal M, Breddam K. 1993. Mutational replacements of the amino acid residues forming the hydrophobic S4 binding pocket of subtilisin 309 from *Bacillus lentus*. *Biochemistry* 32:8994-8999.
- Sprang S, Standing T, Fletterick RJ, Stroud RM, Finer-Moore J, Xuong NH, Hamlin R, Rutter WJ, Craik CS. 1987. The three-dimensional structure of Asn 102 mutant of trypsin: Role of Asp 102 in serine protease catalysis. *Science* 237:905-909.
- Stein RL, Strimpler AM, Hori H, Powers JC. 1987. Catalysis by human leukocyte elastase: Mechanistic insights into specificity requirements. *Biochemistry* 26:1301-1305.
- Steitz TA, Shulman RG. 1982. Crystallographic and NMR studies of the serine proteases. *Annu Rev Biophys Bioeng* 11:419-444.
- Stroud RM. 1974. A family of protein-cutting proteins. *Sci Am* 23:74-88.
- Svendsen I, Jensen MR, Breddam K. 1991. The primary structure of the glutamic acid-specific protease of *Streptomyces griseus*. *FEBS Lett* 292:165-167.
- Takeuchi Y, Noguchi S, Satow Y, Kojima S, Kumagai I, Miura K, Nakamura KT, Mitsui Y. 1991a. Molecular recognition at the active site of subtilisin BPN': Crystallographic studies using genetically engineered proteinaceous inhibitor SSI (*Streptomyces* subtilisin inhibitor). *Protein Eng* 4:501-508.
- Takeuchi Y, Satow Y, Nakamura KT, Mitsui Y. 1991b. Refined crystal structure of the complex of subtilisin BPN' and *Streptomyces* subtilisin inhibitor at 1.8 Å resolution. *J Mol Biol* 221:309-325.
- Tepljakov AV, van der Laan JM, Lammers AA, Kelders H, Kalk KH, Misset O, Mulleners LSJM, Dijkstra BW. 1992. Protein engineering of the high-alkaline serine protease PB92 from *Bacillus alcalophilus*: Functional and structural consequences of mutation at the S4 substrate binding pocket. *Protein Eng* 5:413-420.
- Thompson RC, Blout ER. 1970. Dependence of the kinetic parameters for elastase-catalyzed amide hydrolysis on the length of peptide substrates. *Proc Natl Acad Sci USA* 67:1734-1743.
- Tsu CA, Perona JJ, Schellenberger V, Turck CW, Craik CS. 1994. The substrate specificity of *Uca pugnator* collagenolytic serine protease 1 correlates with the bovine type I collagen cleavage sites. *J Biol Chem* 269:19565-19572.
- Van den Ouweland AMW, Van Duijnhoven HLP, Keizer GD, Dorssers LCJ, Van de Ven WJM. 1990. Structural homology between the human fur gene product and the subtilisin-like protease encoded by yeast KEX2. *Nucleic Acids Res* 18:664-674.
- van der Laan JM, Tepljakov AV, Kelders H, Kalk KH, Misset O, Mulleners LSJM, Dijkstra BW. 1992. Crystal structure of the high-alkaline serine protease PB92 from *Bacillus alcalophilus*. *Protein Eng* 5:405-411.
- Van de Ven WJ, Roebroek AJ, Van Duijnhoven HL. 1993. Structure and function of eukaryotic proprotein processing enzymes of the subtilisin family of serine proteases. *Crit Rev Oncogen* 4:115-136.
- Warshel A, Naray-Szabo G, Sussman F, Hwang JK. 1989. How do serine proteases really work? *Biochemistry* 28:3629-3637.
- Watson HC, Shotton DM, Cox JC, Muirhead H. 1970. Three-dimensional Fourier synthesis of tosyl elastase at 3.5 Å resolution. *Nature* 225:806-811.

- Wei AZ, Mayr I, Bode W. 1988. The refined 2.3 Å crystal structure of human leukocyte elastase in a complex with a valine chloromethyl ketone inhibitor. *FEBS Lett* 234:367-373.
- Wells JA, Cunningham BC, Graycar TP, Estell DA. 1986. Importance of hydrogen bond formation in stabilizing the transition state of subtilisin. *Philos Trans R Soc Lond A* 317:415-423.
- Wells JA, Cunningham BC, Graycar TP, Estell DA. 1987a. Recruitment of substrate-specificity properties from one enzyme into a related one by protein engineering. *Proc Natl Acad Sci USA* 84:5167-5171.
- Wells JA, Cunningham BC, Graycar TP, Estell DA, Carter P. 1987b. On the evolution of specificity and catalysis in subtilisin. *Cold Spring Harbor Symp Quant Biol* 52:647-652.
- Wells JA, Estell DA. 1988. Subtilisin - An enzyme designed to be engineered. *Trends Biochem Sci* 13:291-297.
- Wells JA, Powers DB, Bott RR, Graycar TP, Estell DA. 1987c. Designing substrate specificity by protein engineering of electrostatic interactions. *Proc Natl Acad Sci USA* 84:1219-1223.
- Wilke ME, Higaki JN, Craik CS, Fletterick RJ. 1991. Crystallographic analysis of trypsin G226A. A specificity pocket mutant of rat trypsin with altered binding and catalysis. *J Mol Biol* 219:525-532.
- Willett WS, Gillmor S, Perona JJ, Fletterick RJ, Craik CS. 1995. Engineered metal regulation of trypsin substrate specificity. *Biochemistry*. Forthcoming.
- Wilson C, Mace J, Agard DA. 1991. Computational method for the design of enzymes with altered substrate specificity. *J Mol Biol* 220:495-506.
- Woodbury RG, Everitt MT, Sanada Y, Katunuma N, Lagunoff D, Neurath H. 1978a. A major serine protease in rat skeletal muscle: Evidence for its mast cell origin. *Proc Natl Acad Sci USA* 75:5311-5313.
- Woodbury RG, Gruzinski GM, Lagunoff D. 1978b. Immunofluorescent localization of a serine protease in rat small intestine. *Proc Natl Acad Sci USA* 75:2785-2789.
- Wright CS, Alden RA, Kraut J. 1969. Structure of subtilisin BPN' at 2.5 Å resolution. *Nature* 221:235-242.
- Yoshida N, Everitt MT, Neurath H, Woodbury RG, Powers JC. 1980. Substrate specificity of two chymotrypsin-like proteases from rat mast cells. Studies with peptide 4-nitroanilides and comparison with cathepsin G. *Biochemistry* 19:5799-5804.
- Zerner B, Bender ML. 1964. The kinetic consequences of the acyl-enzyme mechanism for the reactions of specific substrates with chymotrypsin. *J Am Chem Soc* 86:3669-3674.
- Zhou GW, Guo J, Huang W, Fletterick RJ, Scanlan TS. 1994. The three-dimensional structure of a catalytic antibody with active site similarity to serine proteases. *Science* 265:1059-1064.

Exhibit 31

24745

GENOMICS 44, 309–320 (1997)
ARTICLE NO. GE974845

Cloning of the TMPRSS2 Gene, Which Encodes a Novel Serine Protease with Transmembrane, LDLRA, and SRCR Domains and Maps to 21q22.3

Ariane Paoloni-Giacobino,* Haiming Chen,* Manuel C. Peitsch,†
Colette Rossier,* and Stylianos E. Antonarakis*‡,1

*Laboratory of Human Molecular Genetics, Department of Genetics and Microbiology, Geneva University Medical School, Geneva; †Glaxo Institute for Molecular Biology, Geneva; and ‡Division of Medical Genetics, Cantonal Hospital of Geneva, 1211 Geneva, Switzerland

Received March 24, 1997; accepted June 6, 1997

To contribute to the development of the transcription map of human chromosome 21 (HC21), we have used exon trapping from pools of HC21-specific cosmids. Using selected trapped exons, we have identified a novel gene (named TMPRSS2) that encodes a multimeric protein with a serine protease domain. The TMPRSS2 3.8-kb mRNA is expressed strongly in small intestine and weakly in several other tissues. The full-length cDNA encodes a predicted protein of 492 amino acids that contains the following domains: (i) A serine protease domain (aa 255–492) of the S1 family that probably cleaves at Arg or Lys residues. (ii) An SRCR (scavenger receptor cysteine-rich) domain (aa 149–242) of group A (6 conserved Cys). This type of domain is involved in the binding to other cell surface or extracellular molecules. (iii) An LDLRA (LDL receptor class A) domain (aa 113–148). This type of domain forms a binding site for calcium. (iv) A predicted transmembrane domain (aa 84–106). No typical signal peptide was recognized. The gene was mapped to 21q22.3 between markers ERG and D21S56 in the same PI as MX1. The physiological role of TMPRSS2 and its involvement in trisomy 21 phenotypes or monogenic disorders that map to HC21 are unknown. © 1997 Academic Press

INTRODUCTION

Human chromosome 21 (HC21) is the smallest chromosome, with a long arm (21q) of around 40 Mb, containing approximately 600–1000 genes (reviewed in Antonarakis, 1993), and a short arm (21p) of around 10–15 Mb, which

is highly homologous to those of the other four human acrocentric chromosomes. To date, about 75 HC21 genes have been cloned and partially characterised [Genome DataBase, <http://gdbwww.gdb.org>, and SWISS-PROT, <http://www.expasy.ch>]. Trisomy for human chromosome 21 is the most common chromosomal abnormality at birth, leading to the phenotypes known as Down syndrome (Epstein, 1989). In addition, the loci for several monogenic disorders have been mapped to HC21. Dense linkage maps and almost complete physical maps of 21q have already been obtained and are now extensively used for the characterization of HC21 genes and the efforts to determine the nucleotide sequence of HC21. The cloning and characterization of HC21 genes are a necessary step for the understanding of Down syndrome and the molecular etiology of monogenic disorders mapping on this chromosome.

In our laboratory, systematic exon-trapping experiments have been performed to identify portions of HC21 genes, clone and characterize the corresponding full-length cDNAs and genes, and participate in the international effort to create a transcription map of HC21 (Cheng *et al.*, 1994; Peterson *et al.*, 1994; Tassone *et al.*, 1994; Lucente *et al.*, 1995; Chen *et al.*, 1996). We report here the cloning of a novel serine protease gene (TMPRSS2), which is expressed mainly in the small intestine, but also in lower levels in several other tissues, and which maps to 21q22.3. The predicted polypeptide of TMPRSS2 also contains a transmembrane domain, a scavenger receptor cysteine-rich (SRCR) domain, and an LDL receptor class A (LDLRA) domain, and it probably belongs to the type II integral membrane proteins. The TMPRSS2 gene is homologous to, but different from, the human enteropeptidase gene, which maps to a different region of HC21 (21q21).

MATERIALS AND METHODS

Exon Trapping

Pools of chromosome 21-specific cosmids from the LL21NCO2 library (kindly supplied by P. de Jong) were used in exon-trapping

Sequence data from this article have been deposited with the GenBank Data Library under Accession Nos. U75329 (cDNA) and X88229, X88228, X88321, X88043, and X88047 (trapped exons).

To whom correspondence should be addressed at Division de Génétique Médicale, Centre Médical Universitaire, 1 rue Michel-Servet, 1211 Genève 4, Switzerland. Telephone: 41-22-7025707. Fax: 41-22-7025706. E-mail: Stylianos.Antonarakis@medecine.unige.ch.

experiments (Buckler *et al.*, 1991; Church *et al.*, 1994; Gibco BRL Manual 18449-017). *EcoRI*- and *PstI*-digested cosmids were subcloned into pSPL3 vector, and plasmid DNA was used to transfect Cos7 mammalian cells using lipofectACE (Gibco BRL). Total RNA was isolated from Cos7 cells 24 h after transfection, cDNA was synthesized, and PCR products were subcloned into pAMP10 vector by UDG (uracil DNA glycosylase) cloning. After elimination of cryptically spliced, pSPL3-derived clones by oligonucleotide screening, the inserts of individual pAMP10 clones were subjected to nucleotide sequencing on an ABI373A automated sequencer by dideoxy terminator fluorescence method using *Taq* polymerase. Nucleic acid and amino acid homologies of the resulting sequences were analyzed through BLASTN and BLASTX searches of the nonredundant database (Altschul *et al.*, 1990).

Cloning of TMPRSS2 cDNA

The 216-bp PCR product derived from trapped exon HMC26A01 with oligonucleotide primers (26A01A, 5'-GCCTGCGGGTCAAC-TTGAAC-3', and 26A01B, 5'-GGCGGCTGTACGATCCACTC-3') was used as a probe to screen approximately 500,000 clones of a human heart λ gt10 cDNA library (Clontech HL3026a). One positive clone (APG1) was isolated, and the 2.4-kb insert was subcloned into the pAMP10 vector and sequenced in both directions using standard oligonucleotide walking protocols for the ABI373 automated sequencer. The nucleotide sequence was verified using RT-PCR products from intestine poly(A)⁺ mRNA.

Chromosomal Mapping

Two independent methods were used to assign TMPRSS2 to a human chromosome. First, PCR amplification of the trapped exon HMC26A01 with specific oligonucleotide primers (26MAP1, 5'-CAG-GCTTCTGCAGCTTCATC-3', and 26MAP2, 5'-CAATCCATGGCA-TTGGACGG-3') was performed on the genomic DNA from a panel of somatic cell hybrids with defined segments of HC21. Second, the insert of the initial trapped exon HMC26A01 was used to probe high-density filters of cosmids from the HC21-specific LL21NCO2 library. Finally, PCR amplification using either oligonucleotide primers 26MAP1 and 26MAP2 or 26A01A and 26A01B was used on DNAs from a panel of HC21-derived YACs.

5'- and 3'-RACE (Rapid Amplification of cDNA Ends)

To obtain the 5' end of the TMPRSS2 cDNA, 5'-RACE was performed on human small intestine cDNA. From 1 μ g of poly(A)⁺ RNA (Clontech 6547-1) cDNA was made with the Marathon cDNA Amplification kit (K-1802-1), and 5'-RACE using nested PCR primers was carried out with the enzyme *Taq* Expand High Fidelity (Boehringer Mannheim) according to the manufacturer's protocol. The gene-specific primers were 26A01B (see above) and AP26BB (5'-CCGCTG-TCATCCACTATTCC-3'). In two different experiments the same PCR product of 670 bp was generated and subjected to nucleotide sequencing. 3'-RACE was carried out using gene specific primers AP26G (5'-GGTCTGCGCTGTGCCAAGC-3') and AP26K (5'-GTC-TGGCTTTGGCACTCTCTGC-3'), and a PCR product of approximately 2.0 kb was generated.

Northern Blot Analysis

The cDNA clone APG1 containing the complete coding sequence was used to probe two Northern blots, each containing poly(A)⁺ RNA from eight human adult tissues (Clontech 7759-1, Clontech 7760-1), and one containing four fetal tissues (Clontech 7756-1). Northern Blot analysis was performed using standard protocols, with high-stringency washing. A control hybridization using a human actin probe was used for determination of the amount of RNA loaded in these Northern blots.

Comparative Protein Modeling

The sequences of both LDLRA and protease domains of TMPRSS2 were submitted to the SWISS-MODEL automated comparative pro-

tein modeling server (Peitsch, 1995, 1996). The models were made as follows:

LDLRA domain. SWISS-MODEL could not automatically provide a 3D structure of this domain since the degree of identity with the most similar sequence of known 3D structure was less than 30%. Using BLAST (Altschul *et al.*, 1990), we identified the Brookhaven Protein Data Bank entry 1LDL (NMR structure of the LDLR1 domain) (Daly *et al.*, 1995) as the suitable modeling template. We then aligned the TMPRSS2 LDLRA domain with the sequence of 1LDL and submitted the sequence alignment to SWISS-MODEL using the Optimise mode.

Serine protease domain. This domain was modeled using the First Approach mode of SWISS-MODEL, which provides fully automated template identification and multiple sequence alignment prior to model building. Chymotrypsin (P17538) was identified as a suitable modeling template. The template and TMPRSS2 protease sequences were automatically aligned and the model generation proceeded to the end without human intervention. Sequence to structure fitness analysis using both 3D-1D profiles (Lüthy *et al.*, 1992) and Prosal (Sippl, 1993) did not show any obvious discrepancies. The coordinates of both the LDLRA and the serine protease domain of TMPRSS2 can be found in the SWISS-MODEL Repository (<http://www.expasy.ch/swissmod/swmr-top.html>).

RESULTS

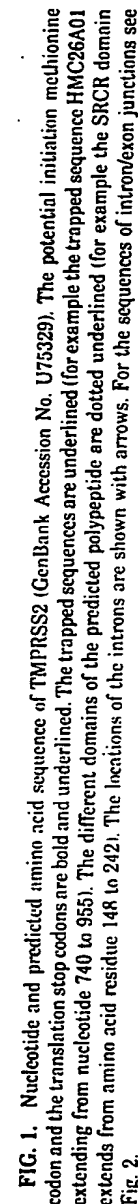
Exon Trapping Identified a Clone with Homology to Human Proteases

To clone partial gene sequences from human chromosome 21 we have used pools of cosmids (from the LL21NCO2-Q library) in an exon-trapping experiment and have identified more than 550 different potential exons (Chen *et al.*, 1996). One trapped sequence HMC26A01 (GenBank X88229) of 216 bp showed a strong homology to a large list of serine proteases from human and other species. BLASTX analysis, for example, revealed a 55% amino acid identity to human prostatic (GenBank L41351; $P = 1.3 \times 10^{-15}$). Other representative homologies included human elastase (P08218), *Erinaceus europaeus* plasminogen (U33171), and pig human coagulation factor IX (P16293). Because this HMC26A01 trapped sequence was probably derived from a undescribed human serine protease, we set out to clone and initially characterize the full-length cDNA of the corresponding human gene.

Isolation of Full-Length TMPRSS2 Coding Sequences

Clone HMC26A01 was used to screen approximately 500,000 clones of a human heart λ gt10 cDNA library (this library was chosen because of the expression pattern in Northern blots; see below). One positive clone (APG1), containing a 2.4-kb-long insert, was obtained, subcloned into the pAMP10 vector, and subjected to nucleotide sequence. 5'-RACE from intestinal mRNA (again chosen because of the expression pattern) using oligonucleotides close to the 5' end of the APG1 clone extended the 5'UTR sequence by about 150 nucleotides. Sequence analysis from both strands revealed an open reading frame of 492 amino acids starting from the most N-terminal methionine codon. The 3'UTR from the original clone APG1 was approximately 0.95 kb. Figure 1 shows the complete nucleotide

1080
1010
1020
1030
1040
1050
1060
1070
1080
1090
1100
1110
1120
1130
1140
1150
1160
1170
1180
1190
1200
1210
1220
1230
1240
1250
1260
1270
1280
1290
1300
1310
1320
1330
1340
1350
1360
1370
1380
1390
1400
1410
1420
1430
1440
1450
1460
1470
1480
1490
1500
1510
1520
1530
1540
1550
1560
1570
1580
1590
1600
1610
1620
1630
1640
1650
1660
1670
1680
1690
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030
2040
2050
2060
2070
2080
2090
2100
2110
2120
2130
2140
2150
2160
2170
2180
2190
2200
2210
2220
2230
2240
2250
2260
2270
2280
2290
2300
2310
2320
2330
2340
2350
2360
2370
2380
2390
2400
2410
2420
2430
2440
2450
2460
2470
2480
2490
2500
2510
2520
2530
2540
2550
2560
2570
2580
2590
2600
2610
2620
2630
2640
2650
2660
2670
2680
2690
2700
2710
2720
2730
2740
2750
2760
2770
2780
2790
2800
2810
2820
2830
2840
2850
2860
2870
2880
2890
2900
2910
2920
2930
2940
2950
2960
2970
2980
2990
3000
3010
3020
3030
3040
3050
3060
3070
3080
3090
3100
3110
3120
3130
3140
3150
3160
3170
3180
3190
3200
3210
3220
3230
3240
3250
3260
3270
3280
3290
3300
3310
3320
3330
3340
3350
3360
3370
3380
3390
3400
3410
3420
3430
3440
3450
3460
3470
3480
3490
3500
3510
3520
3530
3540
3550
3560
3570
3580
3590
3600
3610
3620
3630
3640
3650
3660
3670
3680
3690
3700
3710
3720
3730
3740
3750
3760
3770
3780
3790
3800
3810
3820
3830
3840
3850
3860
3870
3880
3890
3900
3910
3920
3930
3940
3950
3960
3970
3980
3990
4000
4010
4020
4030
4040
4050
4060
4070
4080
4090
4100
4110
4120
4130
4140
4150
4160
4170
4180
4190
4200
4210
4220
4230
4240
4250
4260
4270
4280
4290
4300
4310
4320
4330
4340
4350
4360
4370
4380
4390
4400
4410
4420
4430
4440
4450
4460
4470
4480
4490
4500
4510
4520
4530
4540
4550
4560
4570
4580
4590
4600
4610
4620
4630
4640
4650
4660
4670
4680
4690
4700
4710
4720
4730
4740
4750
4760
4770
4780
4790
4800
4810
4820
4830
4840
4850
4860
4870
4880
4890
4900
4910
4920
4930
4940
4950
4960
4970
4980
4990
5000
5010
5020
5030
5040
5050
5060
5070
5080
5090
5100
5110
5120
5130
5140
5150
5160
5170
5180
5190
5200
5210
5220
5230
5240
5250
5260
5270
5280
5290
5300
5310
5320
5330
5340
5350
5360
5370
5380
5390
5400
5410
5420
5430
5440
5450
5460
5470
5480
5490
5500
5510
5520
5530
5540
5550
5560
5570
5580
5590
5600
5610
5620
5630
5640
5650
5660
5670
5680
5690
5700
5710
5720
5730
5740
5750
5760
5770
5780
5790
5800
5810
5820
5830
5840
5850
5860
5870
5880
5890
5900
5910
5920
5930
5940
5950
5960
5970
5980
5990
6000
6010
6020
6030
6040
6050
6060
6070
6080
6090
6100
6110
6120
6130
6140
6150
6160
6170
6180
6190
6200
6210
6220
6230
6240
6250
6260
6270
6280
6290
6300
6310
6320
6330
6340
6350
6360
6370
6380
6390
6400
6410
6420
6430
6440
6450
6460
6470
6480
6490
6500
6510
6520
6530
6540
6550
6560
6570
6580
6590
6600
6610
6620
6630
6640
6650
6660
6670
6680
6690
6700
6710
6720
6730
6740
6750
6760
6770
6780
6790
6800
6810
6820
6830
6840
6850
6860
6870
6880
6890
6900
6910
6920
6930
6940
6950
6960
6970
6980
6990
7000
7010
7020
7030
7040
7050
7060
7070
7080
7090
7100
7110
7120
7130
7140
7150
7160
7170
7180
7190
7200
7210
7220
7230
7240
7250
7260
7270
7280
7290
7300
7310
7320
7330
7340
7350
7360
7370
7380
7390
7400
7410
7420
7430
7440
7450
7460
7470
7480
7490
7500
7510
7520
7530
7540
7550
7560
7570
7580
7590
7600
7610
7620
7630
7640
7650
7660
7670
7680
7690
7700
7710
7720
7730
7740
7750
7760
7770
7780
7790
7800
7810
7820
7830
7840
7850
7860
7870
7880
7890
7900
7910
7920
7930
7940
7950
7960
7970
7980
7990
8000
8010
8020
8030
8040
8050
8060
8070
8080
8090
8100
8110
8120
8130
8140
8150
8160
8170
8180
8190
8200
8210
8220
8230
8240
8250
8260
8270
8280
8290
8300
8310
8320
8330
8340
8350
8360
8370
8380
8390
8400
8410
8420
8430
8440
8450
8460
8470
8480
8490
8500
8510
8520
8530
8540
8550
8560
8570
8580
8590
8600
8610
8620
8630
8640
8650
8660
8670
8680
8690
8700
8710
8720
8730
8740
8750
8760
8770
8780
8790
8800
8810
8820
8830
8840
8850
8860
8870
8880
8890
8900
8910
8920
8930
8940
8950
8960
8970
8980
8990
9000
9010
9020
9030
9040
9050
9060
9070
9080
9090
9100
9110
9120
9130
9140
9150
9160
9170
9180
9190
9200
9210
9220
9230
9240
9250
9260
9270
9280
9290
9300
9310
9320
9330
9340
9350
9360
9370
9380
9390
9400
9410
9420
9430
9440
9450
9460
9470
9480
9490
9500
9510
9520
9530
9540
9550
9560
9570
9580
9590
9600
9610
9620
9630
9640
9650
9660
9670
9680
9690
9700
9710
9720
9730
9740
9750
9760
9770
9780
9790
9800
9810
9820
9830
9840
9850
9860
9870
9880
9890
9900
9910
9920
9930
9940
9950
9960
9970
9980
9990
10000



...agggcaccctctctctgtttctctgcaag /TGGGCAGCAA.....AATCGGTGTG/ gkgagtcagccttaacctgggaagggaact...
G110 V149
...aactcatggataatcctccctctctgtag /TTCGCTCTA.....TGGGCTATAA/ gkgagtcagccttaacctgggaagggaact...
R150 K191
...cgtgaccagaatttcccgctctctctgtag /TGATGCCTGT.....TCTTTACGCT/ gkgagtcagccttaacctgggaagggaact...
D229 C241
...ctgagatactgagtcctctctctctccag /ACCTCTTAAC.....ACTTTCAACG/ gkgagtcagccttaacctgggaagggaact...
P301 D359
...ggctcaactgtgtttctctctctgaaacag /ACCTAGTGAA.....GAGGAGAAAG/ gkgagtcagccttaacctgggaagggaact...
L360 G391
...tgggagctcaacaagtcctccctgctcttag /GGAAGACCTC.....TTCTTGCCAG/ gkgagtcagccttaacctgggaagggaact...
K392 Q438
...ctgctctctgtacctgtgtgtctccacag /GGTGACAGTG.....ATGAAGGCAA/ gkgagtcagccttaacctgggaagggaact...
G439 N491
...cactttttttttctctatttgaaacaggag /ACGGCTAATccacatggctctctgctctgacgtcgp(3'UTR)...
G492 *

FIG. 2. Intron/exon junctions of the TMPRSS2 gene as determined by comparison of the cDNA sequence to the publicly available sequences of the human P1 clone 35-H5-C8 (Martin *et al.*, 1994; Genbank Accession Nos. L35675-L35682).

and predicted amino acid sequence of TMPRSS2. This cDNA was verified by RT-PCR amplifications from intestinal RNA using pairs of oligonucleotide primers from the cDNA sequence. Interestingly, no ESTs identical to portions of the TMPRSS2 cDNA sequence were identified in the dbEST database of GenBank (search of February 18, 1997). A number of additional exons from the Chen *et al.* (1996) study were identical to portions of the TMPRSS2 cDNA, including HMC44E11 (GenBank X88043), HMC26A05 (GenBank X88228), HMC19A07 (GenBank X88321), and HMC44D02 (GenBank X88047).

Intron/Exon Junctions

Homology searches with sequences available in the public databases revealed identity of discontinuous regions of the TMPRSS2 cDNA with portions of human P1 clone 35-H5-C8 which was sequenced by Martin and co-workers (Martin *et al.*, 1994; GenBank Accession Nos. L35675-L35682). The comparison of the cDNA sequence of TMPRSS2 with the genomic sequence of human P1 revealed intron/exon junctions that are shown in Fig. 2. Not all such junctions are reported in the figure since the sequence of the entire P1 clone was not available in the public databases. It is likely that there are additional introns 5' to codon 110 and between codons 191 and 229 and codons 241 and 301.

Mapping of TMPRSS2 to Chromosome 21

PCR amplification was performed with oligonucleotide primers 26MAP1 and 26MAP2 on genomic DNA from rodent-human somatic cell hybrids that contained either single human chromosomes (NIGMS 2; Drwina *et al.*, 1993) or specific segments of HC21 (Patterson *et al.*, 1993). The expected 155-bp PCR product was present in somatic cell hybrids WAV17, E7b, 725, 2Fur1, R50-3, GA9-3, 9528C-1, 1881C-13b, 8q-, ACEM 2-10d, JC6A, and 1x4; in contrast, somatic cell hybrids

21q+, 6918-8a1, and MRC2-G did not show amplification (data not shown). These data localized this human protease to the region 21q22.3 between markers ERG and D21S56 (Fig. 3).

We used exon HMC26A01 to probe a subset of the cosmid library LL21NC02. One cosmid, Q20A3, was identified as positive. PCR on this cosmid with the same primers 26MAP1 and 26MAP2 produced the expected 155-bp fragment, confirming that Q20A3 contained this exon of TMPRSS2 gene. Yeast DNA from 79 YAC clones, chosen to cover almost all of HC21 (Chu-

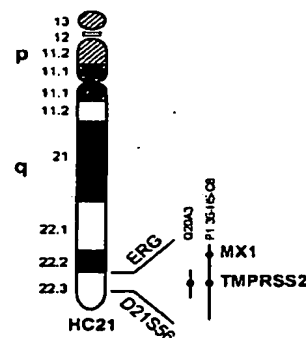


FIG. 3. Schematic representation of the mapping position of the TMPRSS2 gene on chromosome 21 as resulted from PCR amplification of somatic cell hybrids and sequence identities with a chromosome 21 P1 clone (see Results). Representative results from PCR amplification using oligonucleotide primers 26MAP1/26MAP2 (see text) are also shown.

FIG. 1, 7760 while ti
make
with t
26MA
AGC-
3') in
fied. M
of Chu
these
absen
gene s
et al.,
D21S5
clones
deletic
As
TMPR
P1 clo
and co
sion P
gene i
between
sequence
H5-C8
tained
Northe
The
against
from 1
A hybri
ties of

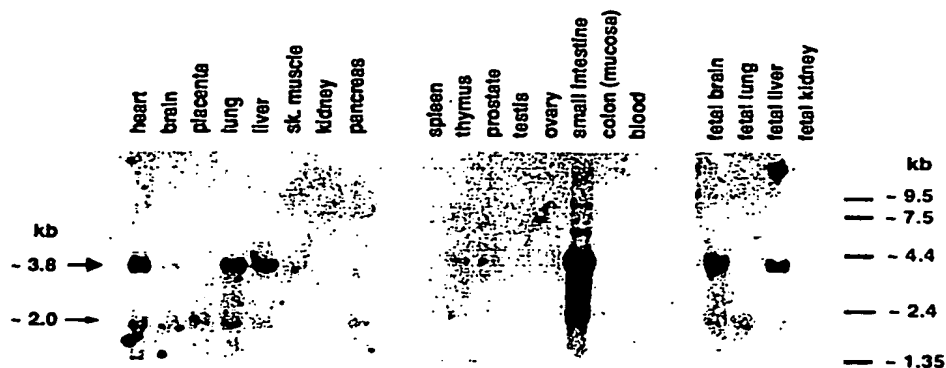


FIG. 4. Northern blot analysis using the TMPRSS2 cDNA as hybridization probe. The RNA filters are from Clontech (Cat. Nos. 7750-1, 7760-1, 7759-1, and 7756-1) and contain 2 μ g of poly(A)⁺ mRNA per tissue indicated. The thick arrow shows the 3.8-kb mRNA species, while the thin arrow depicts the faint 2.0-kb mRNA.

Chumakov *et al.*, 1992), was used for PCR amplification with the two pairs of oligonucleotide primers 26MAP1-26MAP2 and AP26G (5'-GGTTCTGGCTGTGCCAAAGC-3')-AP26H (5'-CCAATGTGCAGGTGGAGACC-3') in the 3'UTR region. No positive YACs were identified. Many single YACs in 21q22.3 from the collection of Chumakov *et al.* (1992) were also tested by PCR with these primers and no amplification was observed. The absence of positive YACs for this human TMPRSS2 gene suggests either that the HC21 contig (Chumakov *et al.*, 1992) in the region between markers ERG and D21S56 contains at least one gap or that the YAC clones available to our laboratory have accumulated deletions.

As described above, discontinuous regions of the TMPRSS2 cDNA were identical to portions of human P1 clone 35-H5-C8, which was sequenced by Martin and co-workers (Martin *et al.*, 1994; GenBank Accession Nos. L35675-L35682). This P1 also contained gene MN1, which maps to 21q22.3 in the interval between ERG and D21S56 (Fig. 3). Therefore, this sequence identity of TMPRSS2 with portions of P1 35-H5-C8 is in agreement with the mapping position obtained using the somatic cell hybrids.

Northern Blot Analysis

The insert of cDNA clone APG1 was used as a probe against three filters containing 2 μ g of poly(A)⁺ RNA from 16 human adult tissues and 4 human fetal tissues. A hybridization signal corresponding to an mRNA species of approximately 3.8 kb was detected (Fig. 4). The

difference between the 2.4-kb cDNA clone APG1 and the 3.8-kb RNA species detected in the Northern blot is probably due to the continuation of the 3'UTR downstream of the end of clone APG1. 3'-RACE from intestinal RNA using oligonucleotides from clone APG1 (oligonucleotide primers AP26G, see above, and AP26K 5'-GTCTGGCTTTGGCACTCTCTGC-3') revealed a PCR product of approximately 2.0 kb, which corresponds to a mRNA length of 3.8 kb, compatible with the results of the Northern blot analyses (data not shown). The highest level of expression was observed in small intestine, but this gene is also expressed in human adult heart, placenta, lung, thymus, and prostate and in fetal brain and liver. Another weakly hybridizing mRNA species of 2.0 kb was also observed in several tissues. This could be due to alternative splicing, utilization of different transcription start sites and polyadenylation signals, overlapping transcripts, or, most likely, cross-hybridizing transcripts with sequence homologies with TMPRSS2. A human actin probe was used to control the amount of RNA loaded (data not shown). The expression of the TMPRSS2 gene appears to be developmentally regulated since there is strong expression in fetal brain but very little expression in adult brain. In addition, in the lung, expression is high in the adult tissue but low in the fetal tissue.

Type II Transmembrane Protein

Protein prediction programs, which predict transmembrane domains, including http://ulrec3.unil.ch/software/TMPRED_form.html (Hofmann and Stoffel,

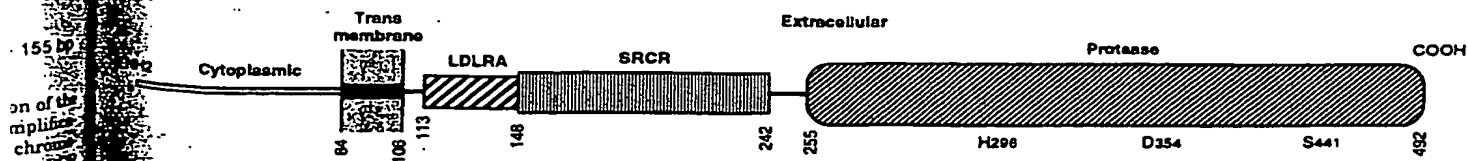


FIG. 5. Schematic representation of the different domains of TMPRSS2. Numbers correspond to codons of the full-length cDNA shown in Fig. 1. For description of the domains see text.

9

[illegible]

FIG. 6. (a) Amino acid sequence comparison of the LDLRA domain of TMPRSS2 with a few selected such domains of other proteins: EK1, 4 (enterokinase—bovine); LDLR1-7 (LDL receptor class A domains—human); L34049a, b (LDL receptor-related protein 2 precursor, megalin—human); U13637a, b (putative vitellogenin receptor precursor—*Drosophila melanogaster*); P07358 (complement C8 β chain precursor—human); U60975 (hybrid receptor gp250 precursor—human); L33417 (VLDL receptor precursor—mouse); and Q99087 (LDL receptor 1 precursor—*Xenopus laevis*). (b) Amino acid sequence comparison of the SLCR domain of TMPRSS2 with a few selected such domains of other proteins: A48231 (cyclophilin C-associated protein precursor—mouse); D13381 (mRNA for macrophage scavenger receptor type 1 subunit—rabbit); P21757 (macrophage scavenger receptor type I and II—human); P21758 (macrophage scavenger receptor type I and II—bovine); P30204 (macrophage scavenger receptor type I and II—mouse); and P16264a-d (IEGG peptide speract receptor precursor). (c) Amino acid sequence comparison of the protease domain of TMPRSS2 with a few other selected proteases: P00766 (chymotrypsinogen—bovine); P03952 (plasma kallikrein precursor—human); P05981 (serine protease hepsin—human); P07146 (trypsinogen precursor—mouse); P07477 (trypsinogen I precursor—human); P07478 (trypsinogen II precursor—human); P14272 (plasma kallikrein precursor—rat); P15157 (α -tryptase precursor—human); P17538 (chymotrypsinogen B precursor—human); P20231 (β -tryptase precursor—human); P20231 (β -tryptase precursor—human); P26262 (plasma kallikrein precursor—mouse); P080773 (enterokinase—human); Q05511 (serine protease hepsin—rat); X07002 (serine protease hepsin—human); X14844 (acrosin precursor—pig); and Y00970 (acrosin precursor—human).

[illegible]

TBP0862
 p00766 (CTPA_BOVH)
 p309332 (KAL_ROMAN)
 p045981 (HBP2_ROMAN)
 p071166 (TPT2_MOUSE8)
 p071677 (TPT2_ROMAN)
 p071708 (TPT2_ROMAN)
 p14272 (KAL_BAT)
 p151357 (TPT2_ROMAN)
 p17538 (CTPA_ROMAN)
 p20231 (TPT2_ROMAN)
 p24262 (KAL_MOUSE8)
 p78073 (HBP2_ROMAN)
 q05511 (HBP2_BAT)
 L14844 (acrosin_P10)
 y009076 (acrosin_ROMAN)

TBP0862
 p00766 (CTPA_BOVH)
 p309353 (KAL_ROMAN)
 p05981 (HBP2_ROMAN)
 p071166 (TPT2_MOUSE8)
 p071677 (TPT2_ROMAN)
 p071708 (TPT2_ROMAN)
 p14272 (KAL_BAT)
 p151357 (TPT2_ROMAN)
 p17538 (CTPA_ROMAN)
 p20231 (TPT2_ROMAN)
 p24262 (KAL_MOUSE8)
 p78073 (HBP2_ROMAN)
 q05511 (HBP2_BAT)
 y009076 (acrosin_P10)
 L14844 (acrosin_ROMAN)

[illegible]

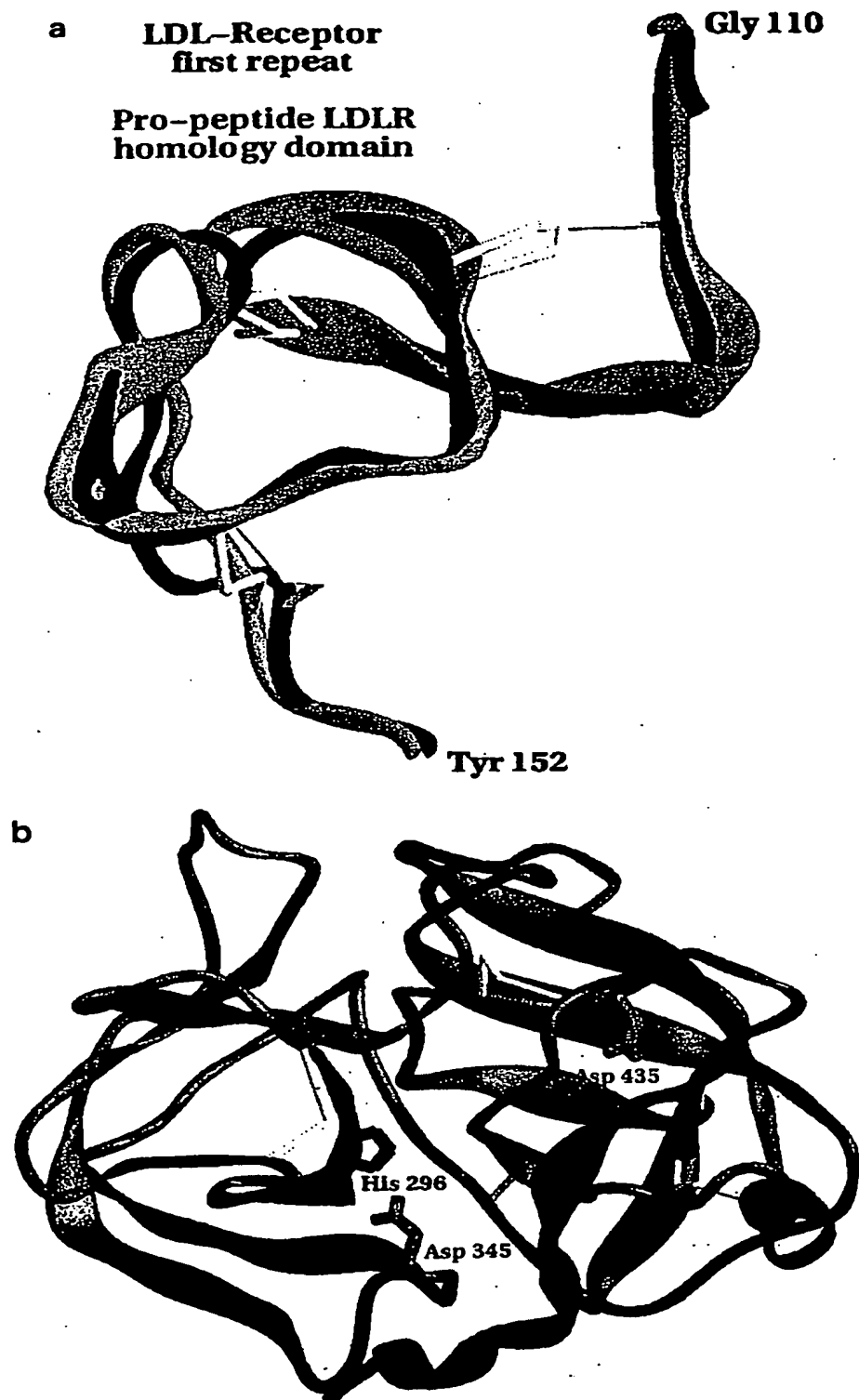
p007166 (CTRA, BOUTN)
 p031592 (LAL, BOUTN)
 p035981 (BEP2, BOUTN)
 p051466 (TWT2, BOUTN)
 p071677 (TWT1, BOUTN)
 p071678 (TWT2, BOUTN)
 p143173 (LAL, BAY)
 p15157 (CTRA, BOUTN)
 p21538 (CTRA, BOUTN)
 p20231 (TWT2, BOUTN)
 p234362 (LAL, BOUTN)
 p045511 (BEP2, BOUTN)
 p045511 (BEP2, BAY)
 x07002 (hep1a, BOUTN)
 L14844 (acrosin, P10)
 x07002 (acrosin, BOUTN)
 p007166 (CTRA, BOUTN)
 p031592 (LAL, BOUTN)
 p035981 (BEP2, BOUTN)
 p051466 (TWT2, BOUTN)
 p071677 (TWT1, BOUTN)
 p071678 (TWT2, BOUTN)
 p143172 (LAL, BAY)
 p15157 (CTRA, BOUTN)
 p21538 (CTRA, BOUTN)
 p20231 (TWT2, BOUTN)
 p234362 (LAL, BOUTN)
 p045511 (BEP2, BAY)
 x07002 (hep1a, BOUTN)
 L14844 (acrosin, P10)
 x07002 (acrosin, BOUTN)

[illegible]

TPR502

FIG. 6—Continued

pig); and Y00970 (acrosin precursor—human).



1993), suggest
were hydrophobic
domain (Fig. 1).
not preceded by
findings are
brane proteins
cytoplasmic
1993). These
polypeptide
similar to the
1988; Tsujie
for cell growth
phology (Kuri
mechanisms

LDLRA Domain

In addition
RSS2 contains
(low-density
from Cys113
motif (PDO)
prod-ent-ry
sity lipoprotei
successive su
LDLRA domain
contains 6 disul
126, 133, 139
have been fo
proteins, inc
sophila puta
kinase comp
C6, C7, C8,
integral men
1995). The ar
domain of T
shown in Fig.
domain and i
of the LDLR
form the bindi
residues betw
important fo
charged sequ
1987; Mahley

The SRCR Domain

An SRCR d
identified in TMF
SRCR domain
and rich in c
derived from
proteins reve
residues (Res
domains are r

FIG. 7. (a) R
while the TMPR
protease domain
His296, blue; As
shown in red.

1993), suggested that amino acids 84–106 of TMPRSS2 were hydrophobic and likely to be a transmembrane domain (Figs. 1 and 5). This hydrophobic sequence is not preceded by a recognizable leader sequence. These findings are compatible with a type II integral membrane protein in which the amino-terminus is at the cytoplasmic side of the membrane (Parks and Lamb, 1993). These features (a type II integral membrane polypeptide with an extracellular protease domain) are similar to those of mammalian hepsins (Leytus *et al.*, 1988; Tsuji *et al.*, 1991). This latter protein is important for cell growth and maintenance of normal cell morphology (Kurachi *et al.*, 1994); however, the underlying mechanisms for the biological activities are unknown.

LDLRA Domain

In addition to the transmembrane domain, TMPRSS2 contains a protein motif of the so-called LDLRA (low-density lipoprotein receptor A) domain extending from Cys113 to Cys148 (Figs. 1 and 5). This structural motif (PDOC00929; <http://www.expasy.ch/cgi-bin/get-prodoc-entry?PDOC00929>) was found in the low-density lipoprotein receptor gene, which contains seven successive such domains (Südhof *et al.*, 1985). A typical LDLRA domain is about 40 amino acids long and contains 6 disulfide-bound cysteines (cysteines 113, 120, 126, 133, 139, and 148 in TMPRSS2). Similar domains have been found in both extracellular and membrane proteins, including the VLDL receptor; gp330; *Drosophila* putative vitellogenin receptor; human enterokinase complement factor I; complement components C6, C7, C8, and C9; perlecan; PKD1; and vertebrate integral membrane protein DGCR2/IDD (Daly *et al.*, 1995). The amino acid comparison of the single LDLRA domain of TMPRSS2 with other similar domains is shown in Fig. 6a. The predicted 3D structure of this domain and its comparison with the first such domain of the LDLR is shown in Fig. 7a. The LDLRA domains form the binding site for LDL and calcium; the acidic residues between the fourth and the sixth cysteines are important for high affinity-binding of positively charged sequences in LDLR ligands (van Driel *et al.*, 1987; Mahley, 1988).

The SRCR Domain

An SRCR domain (Resnick *et al.*, 1994) was also identified in TMPRSS2 extending from Val149 to Leu242. SRCR domains are approximately 100 amino acids long and rich in cysteine. The overall consensus sequence derived from more than 40 such domains from different proteins revealed a consensus sequence at 41 of 101 residues (Resnick *et al.*, 1994). Two groups of SRCR domains are recognized, group A and group B, differing

in the number of conserved cysteines. The SRCR domain of TMPRSS2 contains the pattern compatible with group A SRCR. The sequence homology to different examples of group A SRCR domains is shown in Fig. 6b. The SRCR domains were first found in type I macrophage scavenger receptor (Freeman *et al.*, 1990) but subsequently in many other sequences (for a comprehensive list, see Resnick *et al.*, 1994). The SRCR domain is reminiscent of but different from immunoglobulin domains. Proteins with SRCR domains are either at the cell surface or secreted into plasma or other body fluids. Some proteins such as the WC1 antigen or M130 contain nine or more such domains while others such as the MSR (macrophage scavenger receptor type I) and the secreted CF1 (complement factor 1) or cyclophilin C contain only one domain. The biochemical functions of the SRCR domain have not been established with certainty; however, most of these domains are involved with binding to the cell surface of extracellular molecules.

Protease Domain

The most striking feature of the TMPRSS2 predicted polypeptide is its similarity with members of serine protease family of proteins. The serine protease domain extends from amino acid residue Arg255 to the carboxyl-terminus of the predicted polypeptide. There is approximately 45–55% identity with several members of the serine protease family; the best similarities are with human hepsin (X07002), human enterokinase (P98073), and human kallikrein (P03952). The features of the protease domain of TMPRSS2 are compatible with the S1 family of the SA clan of serine-type peptidases as characterized by Rawlings and Barrett (1994). The prototype of this family is chymotrypsin and the 3D structure of some of its members has already been resolved. For a comprehensive list of the S1 serine-type peptidases see SWISS-PROT (<http://www.expasy.ch/cgi-bin/lists?peptidas.txt>). TMPRSS2 exhibits conservation of serine protease sequence motifs (Fig. 6c); in particular, the active site residues can be identified as His296, Asp345, and Ser441. TMPRSS2 is predicted to cleave after Lys or Arg residues since it contains Asp435 at the base of the specificity pocket (S1 subsite) that binds to the substrate. The predicted 3D structure of the protease domain of TMPRSS2 is shown in Fig. 7b. The protein model was built using the SWISS-MODEL server for automated comparative protein modeling (Peitsch, 1995, 1996) as described under Materials and Methods. It is of interest that TMPRSS2 is highly homologous to hepsin, another protease that contains a transmembrane domain and is thus a type II integral membrane protein with its protease domain

FIG. 7. (a) Ribbon model of the LDLRA domain of TMPRSS2. The NMR structure of the LDL receptor A domain is depicted in blue while the TMPRSS2 LDLRA homology domain is shown in red. The three disulfide bonds are shown in yellow. (b) Ribbon model of the protease domain of TMPRSS2. The full protein structure is depicted as a gray ribbon, while the active sites are shown with colored residues (His296, blue; Asp345, red; Ser441, green). The side chain of Asp435, which determines the Lys/Arg specificity of the TMPRSS2 protease, is shown in red. The three disulfide bonds are depicted in yellow, while two free cysteines are shown as orange bars.

in the extracellular space (Kurachi *et al.*, 1994; Leytus *et al.*, 1988; Tsuji *et al.*, 1991). TMPRSS2 contains nine conserved cysteine residues which by homology to other proteases most likely form the following intrasubunit disulfide bonds Cys826-Cys842, Cys926-Cys993, Cys957-Cys972, and Cys983-Cys1011 and the intersubunit disulfide bond involving Cys758-Cys912 which probably joins the catalytic protease subunit with the nonprotease part of the polypeptide. The protease domain does not contain potential N-glycosylation sites while the remainder of the predicted polypeptide contains two such potential sites (N213, in the SRCR domain, and N249). The amino-terminal Ile of the protease domain is preceded by Arg in the context of a peptide sequence Arg-Ile-Val-Gly-Gly (RIVGG), which is typical for the proteolytic activator site of many serine protease zymogens (Rawlings and Barrett, 1994). The potential cleavage between Arg and Ile, which would be similar to the activation mechanism of other serine protease zymogens, would convert TMPRSS2 to an activated form consisting of a nonprotease and a protease catalytic subunit linked by a disulfide bond that most probably involves Cys758 and Cys912.

DISCUSSION

In this paper we describe the cloning, chromosomal mapping, and initial characterization of a novel gene that maps on human chromosome 21q22.3 and encodes a polypeptide with multiple recognizable domains, namely LDLRA, SRCR, and serine protease domains. In addition, the presence of a transmembrane domain and the absence of a signal peptide suggest that this is a type II integral membrane protein. More biochemical experiments are necessary to further characterize the cellular localization of this protein and its physiological function. The biochemical events for the activation of the probable serine protease activity are unknown but are likely to be similar to those described above. It is of interest that the predicted TMPRSS2 protein contains additional domains (LDLRA and SRCR) that are potentially involved in binding with extracellular molecules or the cell surface. The molecules that are cleaved by or that bind to TMPRSS2 are unknown. There are several tissues that are shown by Northern blot analysis to express the TMPRSS2 gene. The site of the strongest expression is the small intestine; however, other tissues including heart, lung, and liver also showed a significant amount of TMPRSS2 mRNA. The function of this protein in these tissues remains elusive.

Are there any monogenic disorders associated with the TMPRSS2? Several monogenic phenotypes due to mutations in unknown genes have been mapped by linkage analysis to chromosome 21q22.3; these include APECED (Aaltonen *et al.*, 1994; OMIM 240300), an autoimmune disorder, two forms of autosomal recessive deafness (Bonné-Tamir *et al.*, 1996; Veske *et al.*, 1996; OMIM 601072); Knobloch syndrome (Sertie *et al.*, 1996; OMIM 267750); one locus for manic depressive illness (Smyth *et al.*, 1997; OMIM 125480); and one

locus for holoprosencephaly (Muenke *et al.*, 1995; OMIM 236100). All of these phenotypes are mapped more distal to TMPRSS2, and it is therefore unlikely that TMPRSS2 is a candidate gene for any of these disorders.

Many human disorders are due to deficiency of other serine proteases. For example, deficiencies of coagulation factors such as Factor XII (OMIM 234000), Factor X (OMIM 227600), Factor IX (OMIM 306900), and Factor VII (OMIM 227500) belong to these disorders. Additional examples of such disorders are enterokinase deficiency (Hadorn *et al.*, 1969; OMIM 226200), trypsinogen deficiency (Townes, 1965; OMIM 276000), and hereditary pancreatitis due to mutations in the cationic trypsinogen gene (Whitcomb *et al.*, 1996). The generation of mice with targeted disruption of the mouse TMPRSS2 gene will enhance our understanding of the function of this gene and will provide candidate phenotypes for further investigation.

Is the overexpression of three copies of the TMPRSS2 involved in one of the phenotypes of Down syndrome? TMPRSS2 maps outside the so-called Down syndrome critical region (DSCR; between markers D21S17 and ETS2), triplication of which is associated with many phenotypes of Down syndrome (Delabar *et al.*, 1993). However, the existence of a single DSCR has recently been challenged since rare patients with proximal trisomy 21 not including the D21S17-ETS2 region displayed some of the phenotypes of Down syndrome (Korenberg *et al.*, 1994). In addition, a wider region from D21S17 to and including MX1 was associated with several phenotypes, including the heart defect and some dysmorphic features of the syndrome (Delabar *et al.*, 1993; Korenberg *et al.*, 1994). Since the TMPRSS2 gene is within this interval it is formally a candidate for some phenotype(s) of Down syndrome. Transgenic mice that overexpress the murine extracellular protein urokinase-type plasminogen activator have been shown to exhibit abnormal phenotypes (learning disabilities) (Meiri *et al.*, 1994). The study of transgenic mice that overexpress the murine homologue of the human TMPRSS2 gene may contribute to the understanding of the potential involvement of this gene in the pathogenesis of Down syndrome. A mouse model with partial trisomy 16 (which corresponds to a partial human trisomy 21 from APP to MX1) has recently been made (Reeves *et al.*, 1995). It would be of interest to know if the murine homologue of the TMPRSS2 gene is included in the triplicated part of mouse chromosome 16.

ACKNOWLEDGMENTS

We thank P. de Jong for the HC21-specific cosmid library LL2INCO2-Q, D. Patterson for the chromosome-21-specific somatic cell hybrids, and H. S. Scott for critically reading the manuscript. This study was supported by Grant 31-40500.94 from the Swiss FNRS, the European Union Grants GENE-CT93-0015 and PL 970302, and funds from the University and the Cantonal Hospital of Geneva.

CLONING AND MAPPING OF TMRSS2 GENE

319

REFERENCES

- Altonen, J., Bjorses, P., Sandkuijl, L., Perheentupa, J., and Peltola, L. (1994). An autosomal locus causing autoimmune disease: Autoimmune polyglandular disease type I assigned to chromosome 21. *Nature Genet.* 8: 83-87.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
- Antonarakis, S. E. (1993). Human chromosome 21: Genome mapping and exploration, circa 1993. *Trends in Genet.* 9: 142-148.
- Bonné-Tamir, B., DeStefano, A. L., Briggs, C. E., Adair, R., Franklyn, B., Weiss, S., Korostishevsky, M., Frydman, M., Baldwin, C. T., and Farrer, L. A. (1996). Linkage of congenital recessive deafness (gene DFNB10) to chromosome 21q22.3. *Am. J. Hum. Genet.* 58: 1254-1259.
- Buckler, A. J., Chang, D. D., Graw, S. L., Brook, J. D., Haber, D. A., Sharp, P. A., and Housman, D. E. (1991). Exon amplification: A strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. USA* 88: 4005-4009.
- Chen, H., Chrast, R., Rossier, C., Morris, M. A., Lalioti, M. D., and Antonarakis, S. E. (1996). Cloning of 559 potential exons of genes of human chromosome 21 by exon trapping. *Genome Res.* 6: 747-760.
- Cheng, J. F., Boyartchuk, V., and Zhu, Y. W. (1994). Isolation and mapping of human chromosome 21 cDNA: Progress in constructing a chromosome 21 expression map. *Genomics* 23: 75-84.
- Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guasconi, G., Gervy, P., LeGall, I., Soularue, P., Grinas, L., Bougueleret, L., Bellanne-Chantelot, C., Lacroix, B., Barillot, E., Gesnouin, P., Pook, S., Vayseix, G., Frelat, G., Schmitz, A., Sambucy, J. L., Bosch, A., Estivill, X., Weissenbach, J., Vignal, A., Riethman, H., Cox, D., Patterson, D., Gardiner, K., Hattori, M., Sakaki, Y., Ichikawa, H., Ohki, M., Le Paslier, D., Heilig, R., Antonarakis, S. E., and Cohen, D. (1992). A continuum of overlapping clones spanning the entire chromosome 21q. *Nature* 359: 380-386.
- Church, D. M., Stotler, C. J., Rutter, J. L., Murrell, J. R., Trofatter, J. A., and Buckler, A. J. (1994). Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.* 6: 98-105.
- Daly, N. L., Scanlon, M. J., Djordjevic, J. T., Kroon, P. A., and Smith, R. (1995). Three-dimensional structure of a cysteine-rich repeat from the low-density lipoprotein receptor. *Proc. Natl. Acad. Sci. USA* 92: 63334-6338.
- Delabar, J. M., Theophile, D., Rahmani, Z., Chettouh, Z., Blouin, J. L., Prieur, M., Noël, B., and Sinet, P. M. (1993). Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *Eur. J. Hum. Genet.* 1: 114-124.
- Drwinga, H. L., Toji, L. H., Kim, C. H., Greene, A. E., and Mulivor, R. A. (1993). NIGMS human/rodent somatic cell hybrid mapping panels 1 and 2. *Genomics* 16: 311-314.
- Epstein, C. J. (1989). Down syndrome, trisomy 21. In "The Metabolic Basis of Inherited Disease" (C. R. Scriver, A. L. Beaudet, W. S. Sly, and D. Valle, Eds.), pp. 291-326, McGraw-Hill, New York.
- Freeman, M., Ashkenas, J., Rees, D. J. G., Kingsley, D. M., Copeland, N. G., Jenkins, N. A., and Krieger, M. (1990). An ancient, highly conserved family of cysteine-rich protein domains revealed by cloning type I and type II-murine macrophage scavenger receptors. *Proc. Natl. Acad. Sci. USA* 87: 8810-8814.
- Hadorn, B., Tarlow, M. J., Lloyd, J. K., and Wolff, O. H. (1969). Intestinal enterokinase deficiency. *Lancet* i: 812-813.
- Hofmann, K., and Stoffel, W. (1993). Tmbase—A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler* 347: 166.
- Korenberg, J. R., Chen, X. N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, D., Dignan, P., Distech, C., Graham, J. M., Huggins, L., McGillivray, B., Miyazaki, K., Ogasawara, N., Park, J. P., Pagon, R., Pueschel, S., Sack, G., Say, B., Schuffenhauer, S., Soukup, S., and Yamanaka, T. (1994). Down syndrome phenotype: The consequences of chromosomal imbalance. *Proc. Natl. Acad. Sci. USA* 91: 4997-5001.
- Kurachi, K., Torres-Rosado, A., and Tsuji, A. (1994). Hepsin. *Methods Enzymol.* 244: 100-114.
- Leytus, S. P., Loeb, K. R., Hagen, S. F., Kurachi, K., and Davie, E. W. (1988). A novel trypsin-like serine protease (Hepsin) with a putative transmembrane domain expressed by human liver and hepatoma cells. *Biochemistry* 27: 1067-1074.
- Lucente, D., Chen, H. M., Shea, D., Samec, S. N., Rutter, M., Chrast, R., Rossier, C., Buckler, A., Antonarakis, S. E., and McCormick, M. K. (1995). Localization of 102 exons to a 2.5 Mb region of chromosome 21 involved in Down syndrome. *Hum. Mol. Genet.* 4: 1305-1311.
- Lüthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* 356: 83-85.
- Mahley, R. W. (1988). Apolipoprotein E: Cholesterol transport protein with expanding role in cell biology. *Science* 240: 622-630.
- Martin, C. H., Bondoc, M. M., Chiang, A., Cloutier, T., Davis, C. A., Ericsson, C. L., Jaklevic, M. A., Kim, R. J., Lee, M. T., Li, M., Mayeda, C. A., Steiert-El Kheir, A., and Palazzolo, M. J. (1994). Sequencing of the MX1 region of human chromosome 21. [Unpublished] [<http://www2.ncbi.nlm.nih.gov/cgi-bin/genbank?L35675>]
- Meiri, N., Masos, T., Rosenblum, K., Miskin, R., and Dudai, Y. (1994). Overexpression of urokinase-type plasminogen activator in transgenic mice is correlated with impaired learning. *Proc. Natl. Acad. Sci. USA* 91: 3196-3200.
- Muenke, M., Bone, L. J., Mitchell, H. F., Hart, I., Walton, K., Hall-Johnson, K., Ippel, E. F., Dietz-Band, J., Kvaloy, K., Fan, C.-M., Tessier-Lavigne, M., and Patterson, D. (1995). Physical mapping of the holoprosencephaly critical region in 21q22.3, exclusion of SIM2 as a candidate gene for holoprosencephaly, and mapping of SIM2 to a region of chromosome 21 important for Down syndrome. *Am. J. Hum. Genet.* 57: 10747-1079.
- Parks, G. D., and Lamb, R. A. (1993). Role of NH₂-terminal positively charged residues in establishing membrane protein topology. *J. Biol. Chem.* 268: 19101-19109.
- Patterson, D., Rahmani, Z., Donaldson, D., Gardiner, K., and Jones, C. (1993). Physical mapping of chromosome 21. *Prog. Clin. Biol. Res.* 384: 33-50.
- Peitsch, M. C. (1995). Protein modelling by e-mail. *Bio/Technology* 13: 658-660.
- Peitsch, M. C. (1996). ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* 24: 274-279.
- Peterson, A., Patil, N., Robbins, C., Wang, L., Cox, D. R., and Myers, R. M. (1994). A transcript map of the Down syndrome critical region on chromosome 21. *Hum. Mol. Genet.* 3: 1735-1742.
- Rawlings, N. D., and Barrett, A. J. (1994). Families of cysteine peptidases. *Methods Enzymol.* 244: 19-61.
- Reeves, R. H., Irving, N. G., Moran, T. H., Wöhn, A., Kitt, C., Sisodia, S. S., Schmidt, C., Bronson, R. T., and Davisson, M. T. (1995). A mouse model for Down syndrome exhibits learning and behaviour deficits. *Nature Genet.* 11: 177-184.
- Resnick, D., Pearson, A., and Krieger, M. (1994). The SRCR superfamily: A family reminiscent of the Ig superfamily. *Trends Biochem. Sci.* 19: 5-8.
- Sertie, A. L., Quimby, M., Moreira, E. S., Murray, J., Zatz, M., Antonarakis, S. E., and Passos-Bueno, M. R. (1996). A gene which causes severe ocular alterations and occipital encephalocele (Knobloch syndrome) is mapped to 21q22.3. *Hum. Mol. Genet.* 5: 843-847.
- Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins Struct. Funct. Genet.* 17: 355-362.
- Smyth, C., Kalsi, G., Curtis, D., Brynjolfsson J., O'Neill, J., Riskin, L., Moloney, E., Murphy, P., Petursson, H., and Gurling, H. (1997). Two-locus admixture linkage analysis of bipolar and unipolar af-

- fective disorder supports the presence of susceptibility loci on chromosomes 11p15 and 21q22. *Genomics* 39: 271-278.
- Südhof, T. C., Goldstein, J. L., Brown, M. S., and Russell, D. W. (1985). The LDL receptor gene: A mosaic of exons shared with different proteins. *Science* 228: 815-822.
- Tassone, F., Xu, N. X., Wade, H., Weissman, S., and Gardiner, K. (1994). High density transcriptional mapping of chromosome 21 by hybridization selection. *Am. J. Hum. Genet.* 55: 272A.
- Townes, P. L. (1965). Trypsinogen deficiency disease. *J. Pediat.* 66: 275-285.
- Tsuji, A., Torres-Rosado, A., Arai, T., Le Beau, M. M., Lemons, R. S., Chou, S. H., and Kurachi, K. (1991). *J. Biol. Chem.* 266: 16948-16953.
- van Driel, I. R., Goldstein, J. L., Südhof, T. C., and Brown, M. S. (1987). First cysteine-rich repeat in ligand-binding domain of low density lipoprotein receptor binds Ca^{2+} and monoclonal antibodies but not lipoproteins. *J. Biol. Chem.* 262: 17443-17449.
- Veske, A., Oehlmann, R., Younus, F., Mohyuddin, A., Müller-Meskens, B., Mehdi, S. Q., and Gal, A. (1996). Autosomal recessive nonsyndromic deafness locus (DFNB8) maps on chromosome 21q22.1 to a large consanguineous kindred from Pakistan. *Hum. Mol. Genet.* 5: 165-168.
- Whitcomb, D. C., Gorry, M. C., Preston, R. A., Furey, W., Soskice, M. J., Ulrich, C. D., Martin, S. P., Gates, L. K., Jr., Amann, S. T., Toskes, P. P., Liddle, R., McGrath, K., Uomo, G., Post, J. C., and Ehrlich, G. D. (1996). Hereditary pancreatitis is caused by mutation in the cationic trypsinogen gene. *Nature Genet.* 14: 141-145.

GENOMICS
ARTICLE NO.

T
*A

The
been lo
this re
gene, v
which
to flant
overlay
artifici
chromo
marker
quence
precise
showe
this re
PACs
were
which
will b
case g

The
pheno
& pre
The t
abnor
patho
ex m
abnor
involi
m d r

Exhibit 32

New assay technologies for high-throughput screening

Lauren Silverman, Robert Campbell and James R Broach*

The use of high-throughput screening for early stage drug discovery imposes several constraints on the format of assays for therapeutic targets of interest. Homogeneous cell-free assays based on energy transfer, fluorescence polarization spectroscopy or fluorescence correlation spectroscopy provide the sensitivity, ease, speed and resistance to interference from test compounds needed to function in a high-throughput screening mode. Similarly, novel cell-based assays are now being adapted for high-throughput screening, providing for *in situ* analysis of a variety of biological targets. Finally, recent advances in assay miniaturization mark a transition to ultra high-throughput screening, ensuring that identification of lead compounds will not be the rate-limiting step in finding new drugs.

Addresses

Cadus Pharmaceutical Corporation, 777 Old Saw Mill River Road, Tarrytown, NY 10591-6705, USA

*Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA; e-mail: jbroach@molecular.princeton.edu

Current Opinion in Chemical Biology 1998, 2:397-403

<http://biomednet.com/elecref/1367593100200397>

© Current Biology Ltd ISSN 1367-5931

Abbreviations

CRE	cAMP response element
FCS	fluorescence correlation spectroscopy
GFP	green fluorescent protein
HTS	high-throughput screening

Introduction

Continuing advances in molecular biology, human genetics and genomics have accelerated identification of the mechanisms underlying a growing number of human diseases. This progress has increased the number of novel protein targets available for potential therapeutic intervention by drug treatment. Concurrently, novel approaches in combinatorial chemistry and expanded collections of natural products have dramatically increased the number of compounds that can be tested for activity against these targets. The confluence of these two trends towards more potential targets and larger chemical libraries has greatly stimulated adoption of high-throughput screening (HTS) as the primary tool for early stage drug discovery.

HTS is the process by which large numbers of compounds are tested, in an automated fashion, for activity as inhibitors or activators of a particular biological target, such as a cell surface receptor or a metabolic enzyme. Although any assay performed on the bench top can, in theory, be applied in HTS, conversion to an automated format imposes certain constraints that affect the design of the assay in practice. Procedures that are routine at the bench

are often extremely difficult to automate. Also, the more steps required for an assay, the more difficult to automate the HTS. The ideal assay is one that can be performed in a single well with no other manipulation other than addition of the sample to be tested.

A number of assay formats have been developed or modified over the past few years to conform to the constraints imposed by HTS. These assay protocols can be divided into two groups: cell-free assays that measure the biological activity of a relatively pure protein target and cell-based assays that assess the activity of a target, protein by monitoring a biological response of a cell in which the target protein resides. In either case, the protocols require minimal manipulations, can be performed robotically in relatively small volumes, yield robust responses and are relatively impervious to perturbation by solvents and compounds used in drug screening. In this review we describe several of the more recently developed or exploited assay protocols for HTS.

Cell-free assays

The primary goal in adapting cell-free assays to HTS is to minimize the number of steps required in setting up the assay and in detecting the activity, be it an enzymatic reaction or the binding of two components. This goal has been met to a large extent by development of detection systems that do not require separation of the product of the reaction from substrate, or from other components of the assay mixture. Earlier approaches to such homogeneous assay formats relied on proximity-dependent energy transfer. The output of such assays derived from the signal enhancement generated by bringing a source and a distance-dependent amplifier close together. For example, the β -particles of a low-energy radionuclide attached to a ligand will stimulate the fluorescent emission of a scintillant in a bead to which the ligand's receptor is attached [1,2]. More recently, this detection method has been applied to enzymatic reactions, such as that catalyzed by topoisomerase I [3]. As another example of energy transfer assay formats, the rare earth metal lanthanide, Eu^{2+} , when irradiated by light, can transfer its excitation energy in a nonradiative process to the fluorescent protein, allophycocyanin, if the two are in close proximity. This can occur when a Eu^{2+} -derivitized ligand binds to an allophycocyanin-linked receptor [4,5] or a Eu^{2+} -derivitized anti-phosphotyrosine antibody binds to a detector-linked phosphorylated substrate of a tyrosine kinase such as src [6]. Use of time resolved fluorescent procedures assessing emission at specific times following excitation enhances the sensitivity of this technique by reducing interference from background fluorescence, from test compounds or from assay components [6*,7*]. Finally, enzymatic assays suitable for HTS and based on fluorescent resonant energy

transfer between two different forms of green fluorescent protein (GFP) have recently been described [8*].

A number of investigators have exploited fluorescence polarization spectroscopy (FPS) as the basis for homogeneous HTS assays of both enzymatic and binding reactions. When fluorescent molecules in solution are excited with polarized light, the degree to which the emitted light retains polarization depends on the extent to which the fluorescent molecule rotates during the interval between excitation and emission. The rapid rotation of small fluorescent molecules in solution results in substantial loss of polarization. If such small molecules bind to larger molecules, their rotational diffusion is reduced and the retention of polarization is correspondingly increased. Thus, by measuring the relative intensity of emitted light in the planes normal and orthogonal to the plane of the incident polarized light, the extent of rotation of a target molecule, and inferentially, the extent of binding of the target molecule to a larger component, can be calculated. For instance, fluorescent polarization has been used to detect the presence of specific drugs or hormones [9,10], to assess antibody binding of fluorescein-conjugated peptides [11] or to monitor DNA:DNA hybrid formation [12]. The recent availability of a 96-well plate reader [13] with a high sensitivity to fluorescein and fluorescein conjugates has allowed development of 96-well based fluorescent polarization assays. Such high-throughput assays for src family tyrosine kinase activity [14*], for binding of phosphopeptides to Src SH2 domains [15*], for interaction between STAT1 and an γ -interferon receptor-derived phosphotyrosine-containing peptide [16*] and for specific protease activities [17,18*] have recently been described. The sensitivity of fluorescence polarization, the ease and speed with which such assays can be run and the resistance of such assays to interference from absorptive compounds commonly present in complex mixtures [18*] make this procedure highly amenable to HTS.

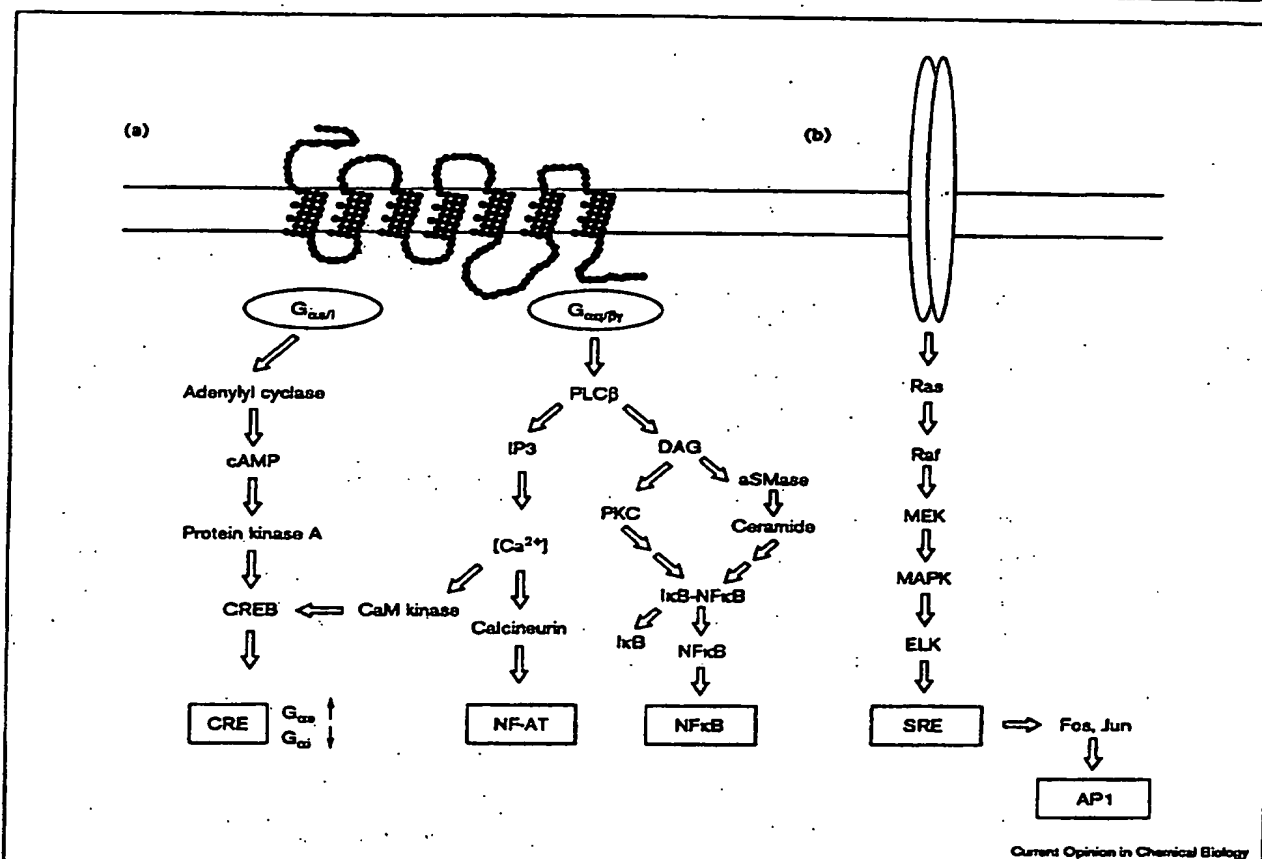
Fluorescence correlation spectroscopy (FCS) represents another recently developed detection format eminently suitable for HTS. FCS measures differences in physical states of a target molecule, such as bound versus free or cleaved versus intact, in a homogeneous mixture [19]. Specifically, FCS measures the burst of fluorescent emission of a molecule passing through a small volume of space, which is defined by a sharply focused laser beam. Small molecules diffuse through the volume rapidly and thus yield short bursts of light. Binding of these small molecules to larger molecules reduces their translational diffusion and correspondingly increases the duration of the bursts of light. Deconvolution of the emission patterns in a sample by appropriate software can yield the relative amount of the bound and unbound states of a fluorescently tagged ligand. This technology can therefore readily be applied to measure receptor-ligand interactions, DNA-protein interactions, nucleic acid hybrid formation and certain enzymatic reactions [20].

Cell-based assays

Cell-based assays are an increasingly attractive alternative to *in vitro* biochemical assays for HTS. Such *in vivo* assays require an ability to examine a specific cellular process and a means to measure its output. For instance, agonist activation of a cell surface receptor or a ligand-gated ion channel can elicit a change in the transcription pattern of a number of genes. This ligand-induced alteration in transcription can be readily captured by using gene fusions, in which a promoter element responsive to receptor activation is fused to the coding region for an enzyme or protein whose levels can be easily measured. Appreciation of the particular signaling pathway associated with a specific receptor allows identification of the appropriate transcriptional response element required to detect a response. Figure 1 depicts a number of signal transduction pathways, indicating the transcriptional response elements coupled to each pathway. Several reporter genes that generate products that can be adapted to HTS format are available [21,22]. These are listed in Table 1, with references to recent innovations in their use [23*,24,25,26*]. For instance, the recent report of novel fluorescent, cell-permeable substrates for β -lactamase documents the use of β -lactamase to detect receptor activation in single cells, making it an attractive assay system for high density HTS [27**].

While cell-based assays using reporter genes have proved effective as an HTS format, detecting more immediate responses to target protein activation provides several advantages, including shorter duration of the assay and fewer false positives from nonspecific interactions. As indicated in Figure 1, such cellular response dependent on activation of a receptor include elevation of a second messenger (for example, Ca^{2+} , cAMP, inositol triphosphate), phosphorylation of an intermediate signaling protein, or subcellular translocation of a signaling molecule. Recent advances in molecular biology and in instrumentation have made it possible to monitor these events in an automated format. For instance, the recent availability of a 96-well fluorescent imaging plate reader (Molecular Devices, Sunnyvale, California, USA) permits HTS of receptor activation by monitoring Ca^{2+} mobilization of cells preloaded with a fluorescent calcium indicator, such as FLUO-3 (Molecular Probes, Eugene, Oregon, USA). In addition, recombinant cells expressing a calcium-sensitive fluorescent protein, such as aequorin [28*] or a hybrid calmodulin-GFP protein [29*], obviate the need for preloading cells with dyes in order to detect calcium fluxes following stimulation. A separate approach to detecting early events following receptor stimulation involves examining relocation of specific components of the signal transduction machinery. For instance, MAP kinase (Figure 1) relocates from the cytoplasm to the nucleus within minutes following stimulation of an upstream G-protein-coupled receptor [30,31]. Similarly, Barak *et al.* [32*] have shown that recruitment of a β -arrestin-GFP fusion protein to the plasma membrane can be used to monitor activation

Figure 1



Signal transduction pathways commonly used in mammalian cell-based high-throughput assays. (a) Agonist-engaged seven transmembrane receptors are functionally linked to the modulation of several well characterized enhancer/promoter elements, the cAMP response element (CRE), nuclear factor of activated T cells (NF-AT), NFκB, serum response element (SRE) and AP1 (48–49). Upon activation of a G_{αs} coupling receptor, adenylyl cyclase is stimulated, producing increased concentrations of intracellular cAMP, stimulation of protein kinase A, phosphorylation of the CRE binding protein (CREB) and induction of promoters with CRE elements. G_{αi} coupling receptors dampen CRE activity by inhibition of the same signal transduction components. G_{αq} coupling receptors and some βγ pairs stimulate phospholipase C (PLC), and the generation of inositol triphosphate (IP₃) and diacylglycerol (DAG). A transient flux in intracellular calcium promotes induction of calcineurin and NF-AT, as well as calmodulin (CaM)-dependent kinase and CREB. Increased DAG concentrations stimulate protein kinase C (PKC) and endosomal/lysosomal acidic sphingomyelinase (aSMase); while the aSMase pathway is dominant, both induce degradation of the NFκB inhibitor IκB as well as NFκB activation. By a poorly understood mechanism, IκB degradation may also be initiated through the MAPK (mitogen-activated protein kinase) cascade (not shown). (b) Growth factor receptor (depicted by ellipses) activation results in recruitment of Sos (not shown) to the plasma membrane, where it stimulates Ras, which recruits the serine/threonine kinase Raf to the plasma membrane. Once activated, Raf phosphorylates MEK kinase, which phosphorylates and activates MAPK and the transcription factor ELK (Ets-like protein, also known as p82 TCF1 [transcription factor 1]). ELK drives transcription from promoters with SRE elements, leading to synthesis of the transcription factors Fos and Jun, that form a transcription complex capable of activating AP1 sites. Seven transmembrane receptors also stimulate the MAPK pathway through βγ subunits, most probably through phosphoinositide 3-kinase γ (PI3Kγ; not shown).

of a number of different G-protein-coupled receptors. Recent advances in microscopic imaging technology, in conjunction with software permitting automated image recognition, provide a means to capture these events in a high-throughput mode.

Cell-based assays have significant advantages over *in vitro* assays. First, the starting material (the cell) self-replicates, avoiding the investment involved in preparing a purified target, in chemically modifying the target to suit the screen, and so on. Second, the targets and readouts are ex-

Table 1

Reporter genes useful for cell-based high-throughput screening.

Reporter genes (source)	Advantages	Disadvantages	References
β -galactosidase (bacterial)	Well characterized; stable, inexpensive substrates; highly sensitive fluorescent or chemiluminescent substrates available; little interference from test compounds; simple readouts (readily automated)	Endogenous activity (mammalian cells); tetrameric (non-linear response at low concentration)	[23*,50]
Luciferase (firefly)	Dimeric; high specific activity; no endogenous activity (low background)	Requires addition of cofactor (luciferin) and presence of O_2 and ATP	[23*]
Alkaline phosphatase (human placental)	Secreted protein (avoids the need for membrane-permeable substrates); inexpensive colorimetric and highly sensitive luminescent assays available	Endogenous activity in some cell types; optimal at pH 9.8	[24,25]
β -lactamase (bacterial)	Monomeric; highly sensitive fluorogenic substrates described; no endogenous activity	Membrane-permeable fluorescent substrates not readily available	[27**]
GFP (jellyfish)	Monomeric; no substrate needed (no manipulations required for assay); no endogenous activity; multiple forms available	Relatively low specific activity	[26*,51,52]

amined in a biological context that more faithfully mimics the normal physiological situation. Third, cell-based assays can provide insights into bioavailability and cytotoxicity. Mammalian cells are expensive to culture and difficult to propagate in the automated systems used for HTS, however.

An alternative to mammalian cell based assays is to recapitulate the desired human physiological process in a micro-organism such as yeast [33]. For instance, signaling via human G-protein-coupled receptors has been reconstituted in yeast to yield a facile growth response or a reporter gene readout ([34,35]; Klein *et al.*, unpublished data). Similarly, mammalian ion channels have been coupled to growth response in yeast [36]. Also, protein-protein interactions, including RAS-RAF association [37] and tyrosine kinase receptor-ligand binding [38], have been faithfully reproduced using the yeast two-hybrid system. Finally, many mammalian transcription factors operate in yeast, including glucocorticoid receptor [39,40] and the retinoic acid receptor and retinoid X receptor families of receptors [41]. The ease and low cost of growing yeast, their ready genetic manipulation, and their resistance to solvents make yeast an attractive option for cell-based HTS.

Miniaturization

Several factors are fueling efforts to increase the speed of HTS and decrease the volume of individual reactions within an HTS format. Split-bead synthesis (see Note added in proof), or other similar approaches to combinatorial chemistry, dramatically increases the number of compounds that can be produced in a library but do so at the cost of quantity of material. In addition, the limited supply of existing compounds within chemical libraries

of pharmaceutical companies, and the growing number of targets against which such compounds can be tested, motivate a frugal approach to use of those compounds. Finally, the reagent costs associated with HTS, when multiplied by the increasing number of assays per run, are becoming a significant cost of early stage drug discovery.

In response to these exigencies, a number of groups have begun to develop formats for very high density screening using very small assay volumes. One approach involves reducing the well size and increasing the density of the assay plate but retaining the overall assay format used in current 96-well based HTS. Densities of 6500 assays in a 10 cm array have been reported for cell-free enzyme based assays [42*] and for ligand binding in cell based assays [43**]. This approach of miniaturizing existing formats significantly increases the number of assays per plate and the overall throughput of the screen but is intrinsically limited by the physical constraints of delivering small volumes to wells, and of detecting responses in a sensitive and timely manner. Accordingly, novel formats have been developed that eschew the assay format based on wells. One approach uses glass chips containing microchannels in which reagents, target proteins and compounds are herded by electrokinetic flow controlled by electric potentials applied at the ends of the channels [44*]. A related approach attains high-throughput both of chemical synthesis and activity assessment by parallel arrays of three-dimensional channels in which flow is controlled by miniature hydrostatic actuators [45]. These approaches provide significant reduction in the volume of assays and a corresponding savings in reagent costs over conventional HTS [45]. In addition, with further development in parallel processing in multiple chips, the number of assays performed in a given period

of time can increase dramatically. This movement to miniaturization is likely to ensure that the initial stage of drug discovery identification of lead compounds will not be the rate-limiting step in finding new drugs.

Conclusions

The last decade has witnessed the emergence across the pharmaceutical industry of the 96-well-based, robotics-driven, high-throughput screening process as the primary tool for identifying active compounds in the first stage of drug discovery. This program has dictated the format of the assays that are used to assess the activities of targets—enzymes, receptors, transporters and so on—that underlie drug discovery in various therapeutic areas. A number of such formats—resonant energy transfer and fluorescent polarization spectroscopy in cell-based assays—have gained widespread acceptance and growing incorporation into high-throughput screening programs. The growing number of potential therapeutic targets, the increasing number of screenable compounds, the accelerating costs of screening and the increasing pressure to generate more lead compounds in a shorter time all conspire to render even the new approaches inadequate for meeting the anticipated throughput requirements, however. Thus, we are likely to witness a movement towards even greater screening throughput by miniaturization and increased reliance on robotics. Whether a new standard format for screening emerges in the near future, or whether a variety of formats are pursued concurrently remains to be seen. Nonetheless, we can anticipate that the exigencies of drug screening will motivate a continued application of state-of-the-art technologies to the process of high-throughput screening.

Note added in proof

For a reference describing split-bead synthesis, see [53].

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Udenfriend S, Gerber LD, Brink L, Spector S: Scintillation proximity assay: a sensitive and continuous isotropic method for monitoring ligand/receptor and antigen/antibody interactions. *Anal Biochem* 1988, 161:494-500.
 2. Cook N: Scintillation proximity assay: a versatile high-throughput screening technology. *Drug Discov Today* 1998, 1:287.
 3. Lerner C, Saito A: Scintillation proximity assay for human DNA topoisomerase I using recombinant biotinyl-fusion protein produced in baculovirus-infected insect cells. *Anal Biochem* 1998, 240:185-198.
 4. Mathis G: Rare earth cryptates and homogeneous fluorimunoassays with human sera. *Clin Chem* 1993, 39:1953-1959.
 5. Mathis G: Probing molecular interactions with homogeneous techniques based on rare earth cryptates and fluorescence energy transfer. *Clin Chem* 1995, 41:1391-1397.
 6. Braunwalder A, Yarwood D, Sills M, Lipson K: Measurement of the protein tyrosine kinase activity of c-src using time-resolved fluorimetry of europium chelates. *Anal Biochem* 1998, 238:159-164.
 7. Gaarde W, Hunter T, Brady H, Murray B, Goldman M: Development of a nonradioactive, time-resolved fluorescence assay for the measurement of jun N-terminal kinase activity. *J Biomol Screen* 1997, 2:213-223.
 8. Mitra R, Silva C, Youvan D: Fluorescence resonance energy transfer between blue-emitting and red-shifted excitation derivatives of the green fluorescent protein. *Gene* 1998, 173:13-17.
 9. Aucourturier P, Prud'homme JL, Lubochinsky B: Fluorescence polarization immunoassay of estradiol. *Diagn Clin Immunol* 1983, 1:310-314.
 10. Erenin SA, Gallacher G, Lotey H, Smith DS, Landon J: Single-reagent polarization immunoassay of methamphetamine in urine. *Clin Chem* 1987, 33:1903-1908.
 11. Wei A-P, Heron JN: Use of synthetic peptides as tracer antigens in fluorescence polarization immunoassays of high molecular weight analytes. *Anal Chem* 1993, 65:3372-3377.
 12. Murakami A, Nakaura M, Nakatsuji Y, Nagahara S, Tran-Cong Q, Makino K: Fluorescent-labeled oligonucleotide probes: detection of hybrid formation in solution by fluorescence polarization spectroscopy. *Nucl Acids Res* 1991, 19:4097-4102.
 13. Jolley Consulting and Research Inc on the World Wide Web: <http://www.jolley.com/>.
 14. Seethala R, Menzel R: A homogeneous, fluorescence polarization assay for src-family tyrosine kinases. *Anal Biochem* 1997, 253:210-218.
 15. Lynch B, Loiacono K, Tlong C, Adams S, MacNeil I: A fluorescence polarization based Src-SH2 binding assay. *Anal Biochem* 1997, 247:77-82.
 16. Wu P, Brasseur M, Schindler U: A high-throughput STAT binding assay using fluorescence polarization. *Anal Biochem* 1997, 249:29-38.
 17. Schade S, Jolley M, Sarauer B, Simonson L: B DIPY-alpha-casain, a pH-independent protein substrate for protease assays using fluorescence polarization. *Anal Biochem* 1998, 243:1-7.
 18. Levine L, Michener M, Toth M, Holwerda B: Measurement of specific protease activity utilizing fluorescence polarization. *Anal Biochem* 1997, 247:83-88.
- The authors describe an assay method to evaluate the activity of protein tyrosine kinases that uses europium chelate-labeled anti-phosphotyrosine antibodies to detect phosphate transfer to a polymeric substrate coated onto microtiter plate wells. Using time-resolved, dissociation-enhanced fluorescence increased sensitivity and reduced interference from test compounds.
- The authors describe a nonradioactive, high-throughput, time-resolved fluorescence assay for jun kinase (JNK) activity using europium-labeled antibody that is specific for amino-terminally phosphorylated c-jun. The optimized europium-based assay is approximately 15-fold more sensitive than a similar 32P-based JNK assay.
- The authors report fluorescent resonance energy transfer (FRET) between two linked variants of the green fluorescent protein, one of which is fused to the amino terminus of a flexible polypeptide linker containing a Factor X protease cleavage site while the second of which is fused to the carboxy terminus of the polypeptide. Cleavage of the peptide linker with Factor Xa yields a marked decrease in energy transfer, making this a viable homogeneous assay format for proteases.
- The authors describe a homogeneous assay for the src kinase family member, Lck, that consists of a fluoresceinylated peptide substrate for the kinase and an anti-phosphotyrosine antibody.
- The authors describe an assay to detect compounds that interfere with the binding of an SH2 domain to its phosphotyrosine peptide target. The assay consists of the src SH2 domain and a fluoresceinylated, phosphotyrosine-containing hexapeptide to which the src SH2 domain binds. Compounds that interfere with this interaction are detected by a decrease in fluorescence polarization.
- The authors describe an assay to detect compounds that interfere with the interaction between STAT1, a transcription factor that is activated upon gamma-interferon binding to its receptor, and the phosphotyrosine-containing peptide derived from gamma-interferon receptor with which it interacts. Binding is evaluated using fluorescence polarization.
- The authors describe a homogeneous fluorescence polarization assay to measure proteolytic cleavage of the peptide substrate for a protease. The peptide substrate was derivatized by biotinylation of the amino terminus and labeled with a fluorescein derivative at the carboxy terminus. Incubation of this substrate with protease and subsequent addition of avidin produced a polarization signal that was proportional to the relative amounts of cleaved

- and uncleaved substrate. The authors demonstrated that the assay does not suffer from interference due to the presence of light absorptive compounds in the mixture.
19. Eigen M, Rigler R: Sorting single molecules: application to diagnostics and evolutionary biotechnology. *Proc Natl Acad Sci USA* 1994, 91:5740-5747.
 20. Storrer S, Henco K: Fluorescence correlation spectroscopy (FCS)-a highly sensitive method to analyze drug/target interactions. *J Recept Signal Transduct Res* 1997, 17:511-520.
 21. Dhundale A, Goddard C: Reporter assays in the high-throughput screening laboratory: A rapid and robust first look? *J Biomol Screen* 1996, 1:115-118.
 22. Suto CM, Ignar DM: Selection of an optimal reporter gene for cell-based high throughput screening assays. *J Biomol Screen* 1997, 2:7-9.
 23. Martin C, Wight P, Dobretsova A, Bronstein I: Dual luminescence-based reporter gene assay for luciferase and β -galactosidase. *Biotechniques* 1996, 21:520-524.
The authors describe highly sensitive chemiluminescent assays for firefly luciferase and β -galactosidase, which can detect as little as 2 fg of luciferase and 8 fg of β -galactosidase, has a dynamic range over seven orders of magnitude of enzyme concentration and both assays can be performed on the same sample.
 24. Bronstein I, Fortin J, Voyta J, Joo R, Edwards B, Olesen C, Lijam N, Kricka L: Chemiluminescent reporter gene assays: sensitive detection of the GUS and SEAP gene products. *Biotechniques* 1994, 17:172-177.
 25. Bronstein I, Martin C, Fortin J, Olesen C, Voyta J: Chemiluminescence: sensitive detection technology for reporter gene assays. *Clin Chem* 1996, 42:1542-1546.
 26. Misteli T, Spector D: Applications of the green fluorescent protein in cell biology and biotechnology. *Nat Biotechnol* 1997, 15:981-984.
The authors present an overview of some of the major applications of an auto-fluorescent protein, the green fluorescent protein, namely its use in protein tagging, in monitoring gene expression and in a variety of biological screens.
 27. Zlokarnik G, Negulescu P, Knapp T, Mere L, Buress N, Feng L, Whitney M, Roemer K et al: Quantitation of transcription and clonal selection of single living cells with beta-lactamase as reporter. *Science* 1998, 279:84-88.
In this chemical and biological tour de force, the authors describe a substrate for β -lactamase that can be readily loaded into living cells and whose hydrolysis by β -lactamase causes a cell-restricted, fluorescence color shift from green to blue that can be detected by eye. The assay allows analysis of gene expression in individual mammalian cells and enables clonal selection by flow cytometry.
 28. Stables J, Green A, Marshall F, Fraser N, Knight E, Sautel M, Milligan G, Lee M et al: A bioluminescent assay for agonist activity at potentially any G-protein-coupled receptor. *Anal Biochem* 1997, 252:115-128.
The authors report that transient transfection of a chinese hamster ovary cell line with the genes for spoozquorn, G_{12} and any of a number of seven transmembrane receptors resulted in a large, concentration-dependent agonist-mediated luminescent response. The authors suggest that this approach provides a basis for a generic mammalian cell microplate assay for the assessment of agonist action at virtually any G-protein-coupled receptor, including orphan receptors for which the physiological signal transduction mechanism may be unknown.
 29. Miyawaki A, Llopis J, Heim R, McCaffery J, Adams J, Ikura M, Tsien R: Fluorescent indicators for Ca^{2+} based on green fluorescent proteins and calmodulin. *Nature* 1997, 388:882-887.
The authors describe the construction and use of novel fluorescent indicators for intracellular Ca^{2+} , which consist of tandem fusions of a blue- or cyan-emitting mutant of the green fluorescent protein (GFP), calmodulin, the calmodulin-binding peptide M13 and an enhanced green- or yellow-emitting GFP. Binding of Ca^{2+} makes calmodulin wrap around the M13 domain, increasing the fluorescence resonance energy transfer (FRET) between the flanking GFPs. The authors can detect free Ca^{2+} dynamics in the cytosol, nucleus and endoplasmic reticulum of single HeLa cells transfected with complementary DNAs encoding chimeras bearing appropriate localization signals.
 30. Lenormand P, Sardet C, Pages G, L'Allemand G, Brunet A, Pouyssegur J: Growth factors induce nuclear translocation of MAP kinases (p42mapk and p44mapk) but not of their activator MAP kinase kinase (p45mapkk) in fibroblasts. *J Cell Biol* 1993, 122:1079-1088.
 31. Gonzalez FA, Seth A, Raden DL, Bowman DS, Fay FS, Davis RJ: Serum-induced translocation of mitogen activated protein kinase to the cell surface ruffling membrane and the nucleus. *J Cell Biol* 1993, 122:1089-1101.
 32. Barak L, Ferguson S, Zhang J, Caron M: A β -arrestin/green fluorescent protein biosensor for detecting G protein-coupled receptor activation. *J Biol Chem* 1997, 272:27497-27500.
The authors describe using a β -arrestin2/green fluorescent protein conjugate to obtain a real-time and single cell based assay to monitor G protein-coupled receptor (GPCR) activation and GPCR-G-protein-coupled receptor kinase or GPCR-arrestin interactions. They show, by confocal microscopy, that the β -arrestin conjugate translocates to the plasma membrane in response to activation of any one of more than 15 different ligand-activated GPCRs, suggesting that β -arrestin binding of an activated receptor is a convergent step of GPCR signaling.
 33. Klein RD, Geary TO: Recombinant microorganisms as tools for high throughput screening for nonantibiotic compounds. *J Biomol Screen* 1997, 2:41-49.
 34. King K, Dohman HG, Thomer J, Caron MG, Leikowitz RJ: Control of yeast mating signal transduction by a mammalian b2-adrenergic receptor and Gsa subunit. *Science* 1990, 250:121-123.
 35. Price LA, Kojkowiak EM, Hadcock JR, Ozenberger BA, Pausch MH: Functional coupling of a mammalian somatostatin receptor to the yeast pheromone response pathway. *Mol Cell Biol* 1998, 18:6188-6195.
 36. Tang W, Rudrudin A, Yang WP, Shaw SY, Knickerbocker A, Kurtz S: Functional expression of a vertebrate inwardly rectifying K^{+} channel in yeast. *Mol Biol Cell* 1995, 6:1231-1240.
 37. van Aelst L, Barr M, Marcus S, Polverino A, Wigler M: Complex formation between RAS and RAF and other protein kinases. *Proc Natl Acad Sci USA* 1993, 90:6213-6217.
 38. Ozenberger BA, Young KH: Functional interaction of ligands and receptors of the hematopoietic superfamily in yeast. *Mol Endocrinol* 1995, 9:1321-1329.
 39. Schena M, Yamamoto KR: Mammalian glucocorticoid receptor derivatives enhance transcription in yeast. *Science* 1988, 241:241-244.
 40. Kralli A, Bohen SP, Yamamoto KR: LEM1, an ATP-binding-cassette transporter, selectively modulates the biological potency of steroid hormones. *Proc Natl Acad Sci USA* 1995, 92:4701-4705.
 41. Hall BL, Smith-McBride Z, Privalsky ML: Reconstitution of the retinoid X receptor function and combinatorial regulation of other nuclear hormone receptors in the yeast *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 1993, 90:6920-6933.
 42. Schullek I, Butler J, Ni Z, Chen D, Yuan Z: A high-density screening format for encoded combinatorial libraries: assay miniaturization and its application to enzymatic reactions. *Anal Biochem* 1997, 246:20-29.
The authors describe a novel, miniaturized high-throughput screening format for assaying combinatorial libraries generated on beads, in which compounds are photolytically released from beads into a high-density well array (>6,500 wells within a standard 96-well microtiter plate footprint) with well volumes as low as 0.37 μ l. Use of the format to detect inhibitors of a member of the matrix metalloproteinase superfamily is described.
 43. You A, Jackman R, Whitesides G, Schreiber S: A miniaturized arrayed assay format for detecting small molecule-protein interactions in cells. *Chem Biol* 1997, 4:969-975.
The authors describe a miniaturized cell-based technique for the screening of ligands prepared by split-pool synthesis, in which spatially defined droplets with uniform volumes of approximately 50-150 nl are arrayed on plastic devices prepared using a combination of photolithography and polymer molding. Using this microtechnology, approximately 8,500 assays, using either yeast cells or mammalian tissue culture, could be performed within the dimensions of a standard 10 cm petri dish.
 44. Hadd A, Raymond D, Halliwell J, Jacobson S, Ramsey J: Microchip device for performing enzyme assays. *Anal Chem* 1997, 69:3407-3412.
The authors describe a novel, miniaturized enzyme assay format, performed within a microfabricated channel network in which precise concentrations of substrate, enzyme and inhibitor were mixed in nanoliter volumes using electrokinetic flow. Reaction kinetics for β -galactosidase obtained by this format were identical to those obtained by conventional assays but required four orders of magnitude less material to do so.
 45. Rogers M: High-throughput screening: miniaturization of high-throughput fluorescence detection for the determination of a few β -galactosidase molecules. *Drug Discov Today* 1997, 2:308.

46. Eder J: Tumor necrosis factor α and Interleukin 1 signaling: do MAPKK kinases connect it all? *Trends Pharmacol* 1997, 320:319-322.
47. Hill CS, Treisman R: Transcriptional regulation by extracellular signals: mechanism and specificity. *Cell* 1995, 80:199-211.
48. Premack BA, Schall TJ: Chemokine receptors: gateway to inflammation and infection. *Nature Med* 1998, 11:1174-1178.
49. Schutze S, Weigmann K, Machleidt T, Kronke M: TNF-induced activation of NF- κ B. *Immunobiology* 1998, 193:193-203.
50. Craig D, Arriaga E, Banks P, Zhang Y, Renborg A, Paicic M, Dovichi N: Fluorescence-based enzymatic assay by capillary electrophoresis laser-induced. *Anal Biochem* 1998, 226:147-153.
51. Anderson M, Tjoe I, Lorincz M, Parks D, Herzenberg L, Nolan G, Herzenberg L: Simultaneous fluorescence-activated cell sorter analysis of two distinct transcriptional elements within a single cell using engineered green fluorescent proteins. *Proc Natl Acad Sci USA* 1998, 95:8508-8511.
52. Heim R, Tsien R: Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Curr Biol* 1998, 8:178-182.
53. Burbaum JJ, Ohtmeyer MM, Reader JC, Henderson I, Dillard LW, Randle TL, Sigal NH, Chelsky D, Baldwin JJ: A paradigm for drug discovery employing encoded combinatorial libraries. *Proc Natl Acad Sci USA* 1995, 92:6027-6031.

Exhibit 33

24715

High-throughput screening: advances in assay technologies

G Sitta Sittampalam*‡, Steven D Kahl*# and William P Janzen†

Both isotopic and nonisotopic assay methodologies are employed in high-throughput screening for drug discovery. Recent advances in cell-based and *in vitro* biochemical assays will be reviewed, with special emphasis on detection technologies amenable to automated 'mix and read' procedures in high-throughput screening. A major trend is the advent of homogenous assay systems which employ fluorescence resonance energy transfer, fluorescence polarization, and fluorescence correlation spectroscopy. Cell-based assay systems have also become popular in high-throughput screens in which active compounds that directly modulate the disease target are identified. Colorimetric and amperometric methods have also been described recently, but are yet to be adapted widely in high-throughput screens.

Addresses

*Research Technologies and Proteins, Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285, USA

†Sphinx Pharmaceuticals, A Division of Eli Lilly and Company, 4615 University Drive, Durham, North Carolina 27707, USA;

e-mail: janzen@lilly.com

‡e-mail: sitta@lilly.com

#e-mail: skahl@lilly.com

Current Opinion in Chemical Biology 1997, 1:384-391

<http://biomednet.com/elecref/1367593100100384>

© Current Biology Ltd ISSN 1367-5931

Abbreviations

DABCYL	4-(4'-dimethyl-aminobenzeneazo)benzoic acid
EDANS	5-(2'-aminoethyl)aminonaphthalene sulfonic acid
FCS	fluorescence correlation spectroscopy
FLIPR	fluorescence imaging plate reader
FPA	fluorescence polarization assay
FRET	fluorescence resonance energy transfer
HTRF	homogeneous time-resolved fluorescence
HTS	high-throughput screening
RET	resonance energy transfer
SPA	scintillation proximity assay
WGA	wheat germ agglutinin

Introduction

The discovery of pharmaceutical agents with novel structures and potential therapeutic activity is a complex process. It usually begins with intensive studies of the physiological and clinical manifestations of diseases, followed by the identification of relevant genes and/or associated biological targets for therapy. Recent advances in molecular biology and DNA sequencing techniques have made tremendous progress toward sequencing large genomes [1]. It is anticipated that the sequencing of the entire human genome, which consists of ~3000 megabases (over 100,000 genes), will be completed in the early part of the next century. Hence the identification of genes that determine the expression of biological targets associated

with human disease is rapidly advancing, opening new and exciting opportunities for the discovery of life-saving drugs.

Coupled with these advances are developments in combinatorial chemistry, where large and structurally diverse chemical libraries are being generated at an unprecedented rate using parallel synthesis [2]. Innovations in powerful computers, automation and software technology have provided an ideal environment to test hundreds of thousands of compounds for biological activity, identifying active molecules or 'hits' that can rapidly develop into potential drugs or 'leads' with desired therapeutic activity.

High-throughput screening (HTS) is the process of testing a large number of diverse chemical structures against disease targets to identify 'hits'. Excellent introductions and reviews on high-throughput screening (HTS) have been published recently [3**,4*,5,6**]. Briefly, current state-of-the-art HTS operations are highly automated and computerized to handle sample preparation, assay procedures and the subsequent processing of large volumes of data. Each one of these steps requires careful optimization to operate efficiently and screen 100-300,000 compounds in a 2-6 month period. Hence a modern HTS operation is a multidisciplinary field involving analytical chemistry, biology, biochemistry, synthetic chemistry, molecular biology, automation engineering and computer science [5].

Central to the HTS process is an *in vitro* biochemical or cell-based assay using a validated biological target representing a disease state. In this paper, we will focus on current assay technologies that are employed in HTS, with emphasis on their advantages and disadvantages. Developing detection technologies with potential applicability to HTS will also be briefly reviewed.

HTS Instrumentation and capabilities

In general, the instrumentation used in HTS assays should be accurate, reliable and easily amenable to automation. Analytical methods should be robust and reproducible, with stable reagents and signal responses. Signal-to-noise (S/N) ratios should be large enough to generate signal windows [7*] that allow reliable detection of 'hits'. Equally important are assays with 'mix and measure' protocols, which are easier to automate than analytical methods with complex separation steps such as centrifugation, washing and filtration. This is particularly true as the industry moves toward ultra-HTS assays which will screen over 100,000 compounds per day [8]. Another advantage of 'mix and measure' assays is that binding measurements are made under equilibrium conditions (without washing, filtration etc.), and are therefore useful for investigating low affinity interactions [9].

Standard HTS assays are currently run in 96-well microtiter plates in batch formats, since automation and detection instruments have been designed to be compatible with these plates. Combinatorial chemical synthesis can also be carried out in 96-well plates, making these plates a standard platform in nearly all HTS operations. Although assays in plates with 384 wells and (as well as 864- and 1536-wells which use the same plate dimensions) are being tested, assay formats based on these high density plate formats have yet to be widely implemented.

Common therapeutic targets for HTS are enzymes, cell surface receptors, nuclear receptors, ion channels, and signal transduction proteins [3*]. Compounds that interact with these targets are usually identified using *in vitro* biochemical assays; however, cell-based assays using engineered mammalian cell lines are now widely employed in HTS. This is because the ligand interaction occurs in the biological environment of the target, which provides opportunities to simultaneously monitor secondary cellular events such as cytosolic Ca^{2+} mobilization and other G-protein-coupled signaling. In addition, the target need not be purified extensively in order to be compatible with the *in vitro* screening conditions. Cell-based assays also screen simultaneously for the bioavailability of test compounds when intracellular targets such as nuclear receptors are involved. A major disadvantage, however, is the cost and difficulty of producing stable, engineered eukaryotic cell lines. Special techniques, instrumentation, and reagents compatible with cell-based assays have to be developed. Once in place, however, HTS laboratories are able to employ cell-based screens routinely. Detection

technologies available for both types of assays will be reviewed below.

Detection technologies

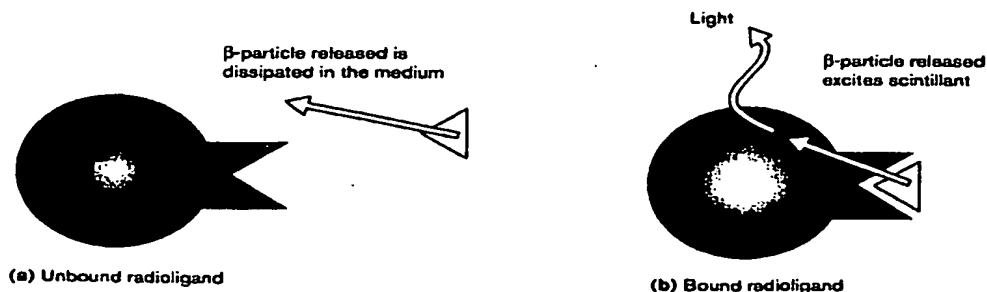
Radiochemical methods

Detection technologies employed in high-throughput screens depend on the type of biochemical pathway being investigated. For example, *in vitro* receptor binding assays with K_d values in the nanomolar to picomolar (nM-pM) range generally employ radiometric detection. The same is true for protein-protein interaction assays with K_d values in the micromolar to nanomolar (μM -nM) range. Enzymatic assays, on the other hand, routinely employ colorimetric, fluorimetric and radiometric detection.

Although filtration-based receptor binding assays have been used extensively in the past (to separate the bound and free radiolabeled ligand), the scintillation proximity assay (SPA) has become the standard assay in many HTS operations, mainly because it does not require a separation step, and can be easily automated [9,10,11*,12*,13,14,15*,16-21]. SPA can also be easily adapted to a variety of enzyme assays [13,14,15*,16] and protein-protein interaction assays [9,18,19].

One version of SPA utilizes polyvinyltoluene (PVT) microspheres or beads ($\sim 5 \mu\text{m}$ diameter, density $\sim 1.05 \text{ g/cm}^3$) into which a scintillant has been incorporated (Figure 1; [8]). When a radiolabeled ligand is captured on the surface of the bead, the radioactive decay occurs in close proximity to the bead, and effectively transfers energy to the scintillant, which results in light emission. When the

Figure 1



Current Opinion in Chemical Biology

Principles of scintillation proximity assay (SPA) technology. (a) The path length of decay for the β -particle released by the isotope is not close enough to the SPA bead and the energy is dissipated in the aqueous medium resulting in little or no detection. (b) When the radioligand is bound to the SPA bead (through a specific capture molecule) the β -particle released is capable of exciting the scintillant contained within the bead and detectable light is emitted.

radiolabel is displaced or inhibited from binding to the bead, it remains free in solution and is too distant from the scintillant for efficient energy transfer. Energy from radioactive decay is dissipated into the solution, which results in no light emission from the beads. Hence the bound and free radiolabel can be detected without the physical separation required in filtration assays.

The outer surface of the SPA bead is coated with a hydrophilic polyhydroxy film that reduces hydrophobicity of the bead to reduce nonspecific interactions. This film has been chemically derivatized to covalently couple generic-capture molecules. PVT beads with the following capture molecules are commercially available: Protein A, avidin, streptavidin, wheat germ agglutinin (WGA), glutathione, and sheep antimouse, donkey antirabbit and donkey antisheep antibodies. All of these capture molecules are used routinely as one member of a detection-pair system. These beads are easily pipetted using automated liquid handling devices into 96-well plates and, therefore, are easily accommodated into HTS operations.

The ideal isotopes for labeling ligands used in SPA assays are ^3H and ^{125}I . This is because the β particles from ^3H have a relatively short pathlength, about $1.5\mu\text{m}$, which easily fulfils the distance requirement for SPA. The Auger electrons emitted by ^{125}I , which travel between approximately $1\mu\text{m}$ and $17.6\mu\text{m}$ in aqueous media, also satisfy this distance requirement. Other commonly used isotopes in biology (^{14}C , ^{35}S , ^{32}P , ^{33}P) emit particles with longer pathlengths and are not suitable for SPA beads, since their decay is detected by the scintillant, even when the ligand is not bound to the surface of the bead (this is called the nonproximity effect). An SPA using ^{33}P -labeled substrate for the cytomegalovirus protease has been reported, however [15]. The decay pathlength for this isotope is $\sim 126\mu\text{m}$, and it is not clear how the nonproximity effect was avoided in this case. In a similar screen using ^{33}P -labeled peptide for calcineurin phosphatase activity, the nonproximity effect was successfully minimized by a simple centrifugation of assay plates [16]. Other enzyme assays for topoisomerase I [13] and *N*-acetylgalactosaminyltransferase [14] utilized ^3H -labeled substrates. The advantage of using ^3H is that the signals can be quite small, and disposal requires special precaution due to its long half-life. Other recent applications of SPA beads include a toxicokinetic study of antisense oligonucleotides in plasma [17] and a kinetic analysis of inositol triphosphate binding to its receptor [20]. It appears that the use of SPA technology may rapidly expand beyond HTS into other areas of drug discovery and development such as genomics, cell metabolism and toxicology.

SPA can also be carried out in scintillating microplates [9,21,22], in which the scintillant is directly incorporated into the plastic, or is coated on the inner surface of the wells. These plates are available from two sources.

Flashplate® is from NENTM Life Science Products (Boston, MA) in which the scintillant is coated on the inner surface of the wells. The Scintistrip® plate is from Wallac-Oy (Turku, Finland) which is made by incorporating the scintillant into the entire plastic. With appropriate washing (not a 'mix and measure' technique) these plates offer the advantage of eliminating nonproximity effects. In addition, these plates are available without licensing fees (required for the bead technology). One example of this is a protein-peptide interaction screen in which the binding of a 13 amino acid phosphopeptide fragment of the epidermal growth factor (EGF) receptor to the GRB2-SH2 binding domain was investigated using the Scintistrip® plates [9]. The screen consisted of adding compounds to be tested and the ^{125}I -labeled phosphopeptide, respectively, to a plate pre-coated with GRB2-SH2 binding domain, followed by a one hour incubation at room temperature. It was, however, necessary to remove all liquid from the wells followed by air-drying the plates before counting. This removal is essential to minimize nonproximity effects which contribute to background noise. An additional advantage of these plates is that they are compatible with other isotopes such as ^{14}C , ^{35}S , ^{33}P , and ^{32}P .

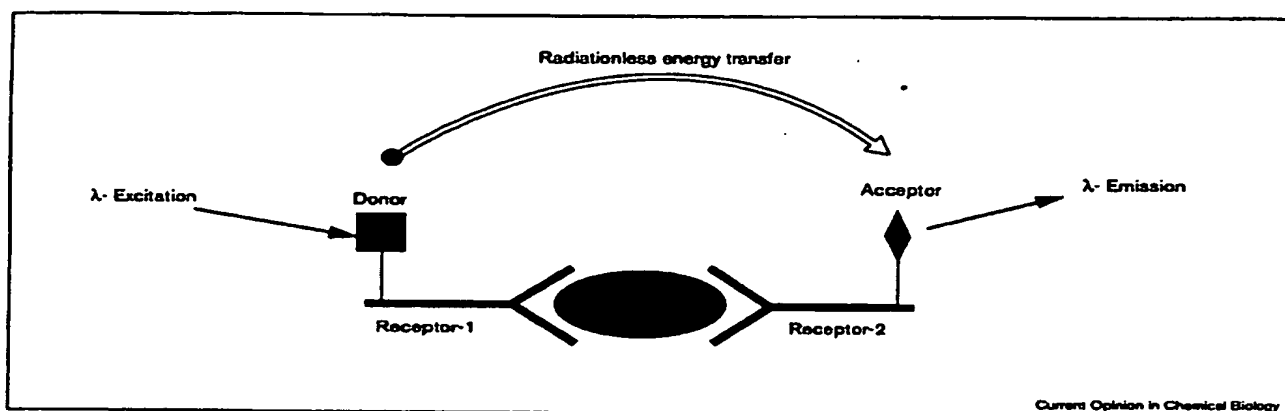
A more recent development is the Cytostar-TM (Amersham Life Sciences, Cardiff, Wales) scintillating microplates [21] which were specially designed for cell-based proximity assays. Scintillant is incorporated into the base plate of microtiter plates and can also detect additional isotopes such as ^{14}C , ^{45}Ca , ^{35}S , ^{33}P . These plates have been successfully used to monitor ^{14}C -labeled thymidine uptake by cultured cells, and to measure $^{45}\text{Ca}^{2+}$ flux through ionotropic glutamate-gated ion channels. The Cytostar-TM plates were also used to detect mRNA transcripts in a high volume *in situ* hybridization [22]. This is an interesting example of how HTS assay concepts are being applied to gene expression and target identification studies.

Non-isotopic detection methods

Colorimetry and luminescence

Colorimetric and luminescence detection methods have significant advantages for HTS laboratories, particularly in light of the cost, safety and disposal issues associated with radiochemical methods. HTS operations require relatively large amounts of reagents during scale-up, operations and follow-up phases. Radiolabeled reagents are expensive, and the scientists running radioactive screens should be adequately trained and monitored. Since luminescence methods can be as sensitive as radioactive methods, with low detection limits, these techniques are being used increasingly in HTS assays [23,24,25-29,30,31-34,35,36,37,38,39,40-42,43,44-51]. Gläzer [24] and Czarnik [125] and the Fluorescent Chemosensors and Biosensors Database on the World Wide Web URL: <http://biomednet.com/fluoro/> have reviewed the utility and need for fluorescence-based

Figure 2



Principles of fluorescence resonance energy transfer. The transfer is inversely proportional to the sixth power of the distance between the donor-acceptor pair, and occurs only when they are in close proximity via binding to the same ligand. Interactions of the labels with the medium and nonspecific fluorescence from the medium itself and the spectral overlap of the donor emission and acceptor absorption can significantly affect the measured signal. Hence the selection of the donor-acceptor pair is critical to the success of energy transfer experiments. Acceptors with long fluorescent lifetimes (microseconds) allow time-resolved measurement of the fluorescence emission. Time-resolved measurements significantly enhance the signal-to-noise ratios, since the fluorescence lifetimes of impurities are generally in the nanosecond time scale.

techniques for biological applications, which can be easily extended to HTS assays.

Resonance energy transfer

Resonance energy transfer (RET; Figure 2) between a fluorophore and chromophore was one of the earliest methods developed for HTS. A peptide substrate for an HIV protease was synthesized with EDANS (at the amino terminus) as the donor fluorophore, and DABCYL (at the carboxyl terminus) as the acceptor chromophore [26]. Energy transfer from EDANS to DABCYL in the intact peptide resulted in quenching of EDANS fluorescence. On cleavage by HIV protease, the fluorescence of the cleaved tetrapeptide-EDANS was restored to the free fluorophore level. Using this assay, inhibitors of HIV protease activity were identified using a simple 'mix and measure' assay format [26]. Although a 40-fold enhancement of the fluorescence signal could be obtained in this assay, there are several disadvantages to the DABCYL-EDANS pair. Many organic and natural product compounds absorb around the absorption and emission maxima of EDANS (λ_{ab} ~ 340 nm, λ_{em} ~ 490 nm). These organic and natural product compounds can also quench the EDANS fluorescence, generating false positives. Any trace contamination of the peptide substrate with free EDANS would result in a high fluorescence background.

Time-resolved fluorescence

A new homogeneous time-resolved fluorescence (HTRF) technology has been described [27]. The assay utilizes fluorescence energy transfer between two fluorophores (a europium cryptate and a 105 kDa phycobiliprotein,

allophycocyanin) as labels. The Eu-trisbipyridine cryptate (TBP-EU³⁺, λ_{ex} ~ 337 nm) has two bipyridyl groups that harvest light and channel it to the caged Eu³⁺. It has a long fluorescence, lifetime and nonradiatively transfers the energy to allophycocyanin when the two labels are in close proximity (> 50% transfer efficiency at a donor-acceptor distance of 9.5 nm). The resulting fluorescence of allophycocyanin (λ_{em} ~ 665 nm) retains the long lifetime of the donor TBP-EU³⁺, allowing time-resolved measurement. Both these labels and their spectroscopic characteristics are very stable in biological media. Several homogeneous *in vitro* biochemical assays based on these two labels have been described [27]: binding of epidermal growth factor (EGF) to its receptor, a Jun/Fos protein-protein interaction and as well as a tyrosine kinase assay. Using this concept, the first HTS assay for a protease enzyme (herpes simplex virus type-1) was recently described by Kolb *et al.* [28].

Cell-based fluorescence assays

The above methodologies are not easily adapted to cell-based assays. An interesting fluorescence resonance energy transfer (FRET) procedure for sensing voltage across cell membranes has been described recently, however [29]. The technique uses membrane permeable, anionic oxonols which rapidly locate on the inner or outer membrane surface depending on polarization state of the membrane. FRET occurs between fluorescein-labeled WGA and the oxonols bound to the outer surface of the membrane at a resting negative potential. At a positive potential, the oxonols are relocated to the inner membrane surface, and the FRET is greatly reduced.

Many fluorescence intensity measurements, including FRET, can be easily configured on a new instrument specifically designed for cell-based HTS assays in 96-well plates called FLIPR [30*]. FLIPR utilizes a water-cooled argon ion laser (5 watt) or a xenon arc lamp and a semiconfocal optical system with a charge-coupled device (CCD) camera to illuminate and image the entire plate. The spatial resolution of the optics is $\sim 200\mu\text{m}$ at the cell plane. The plate chamber temperature can be controlled precisely, and a 96-well pipettor head is integrated into the instrument. These features allow accurate measurements of cellular biochemistry in confluent layers of cells at the bottom of plates. FLIPR software can rapidly quantify transient fluorescence signals in intact cells that are growing attached to the bottom of the well. HTS assays involving intracellular calcium, pH and membrane potential measurements have been designed using this instrument [31].

Fluorescence polarization

Another technique that has gained popularity recently is fluorescence polarization or anisotropy [32–34,35*,36,37,38*]. When fluorescently labeled molecules in solution are illuminated with plane-polarized light, the emitted fluorescence will be in the same plane provided the molecules remain stationary. Since all molecules tumble as a result of collisional motion, depolarization of fluorescence emission occurs. This polarization phenomenon is proportional to the rotational relaxation time (μ) of the molecule, which is defined by the expression $3\eta V/RT$. At constant viscosity (η) and temperature (T) of the solution, polarization is directly proportional to the molecular volume (V) (R is the universal gas constant). Hence changes in molecular volume or molecular weight due to binding interactions can be detected as a change in polarization. For example, the binding of a fluorescently labeled ligand to its receptor will result in significant changes in measured fluorescence polarization values for the ligand. Once again, the measurements can be made in a 'mix and measure' mode without physical separation of the bound and free ligands. The polarization measurements are relatively insensitive to fluctuations in fluorescence intensity when working in solutions with moderate optical intensity.

A fluorescence polarization assay (FPA) for the cytomegalovirus protease using a peptide substrate labeled with biotin and 5-(4,6-dichlorotriazinyl)aminofluorescein was reported recently [35*]. This assay is similar to the SPA assay reported earlier [15*], except that the capture reagent is avidin, and it is added to the enzyme substrate mixture. High polarization values were observed when the enzyme was inhibited and the uncleaved substrate became complexed with avidin. Another HTS utilizing an FPA involved the interaction of fluorescein-labeled peptides containing phosphorylated tyrosine with Src-SH2 domains [38*]. In both cases, a 96-well plate reader (FPM-2, Jolley Consulting and Research, Round Lake Illinois, USA) was used for the HTS. Signal from the

entire plate is read in about three minutes, making 50–100 plates/day assays quite feasible in HTS laboratories.

Fluorescence correlation spectroscopy

Fluorescence correlation spectroscopy (FCS) has been recently described for HTS applications [39*,40,41]. FCS measures time-dependent and spontaneous fluctuations in fluorescence intensities in very small volumes (nanoliters). These fluctuations usually result from Brownian motion associated with chemical reactions, diffusion or the flow of fluorescently labeled molecules. The average fluctuation is proportional to the square root of N , where N is the average number of molecules in the volume. Since Brownian diffusion is directly affected by molecular interactions, FCS is an excellent tool to measure binding interactions [23]. Using powerful lasers and autocorrelation techniques, sensitive measurements (at concentrations of $\sim 10^{-12}\text{M}$) can be made both in solution and in cellular compartments. Access to this technology is limited since this instrumentation for HTS is available only through collaborative agreements on a semiexclusive basis [39*].

Cell-based assay systems for HTS have been thoroughly reviewed, with guidelines for selecting appropriate screening systems [43**]. Assay systems using mammalian and insect cells, as well as yeast and bacterial cells have been described. The most common method for detecting ligand interaction with drug targets expressed in cells is to employ a reporter gene [3**,43**,45,46,49,50]. This involves splicing the transcriptional control elements of a target gene (a gene that controls the biological expression and function of a disease target) with a coding sequence of a reporter gene into a vector. This vector is then transfected into a suitable cell line in order to construct a detection system that responds to modulation of the target. Common examples of reporter genes are enzymes such as chloramphenicol acetyltransferase (AT), alkaline phosphatase (AP), firefly and bacterial luciferases, and β -galactosidase. These enzymes can be detected at very low levels using colorimetric, chemiluminescent or bioluminescent products of specific substrates. The chemistry of chemiluminescent and bioluminescent reactions have been reviewed in detail [46,47].

A new reporter system using the β -lactamase enzyme with a membrane permeable fluorogenic substrate has been cited for cell-based assays [3**]. The advantage is that the enzyme is monomeric and has no endogenous activity in mammalian cells. Since fluorescent substrates are not yet commercially available, this system is yet to be used widely in HTS applications.

Future developments and conclusions

Several new trends can be observed in the recent HTS literature ([52–56,57**,58–69]. The use of 384-well plates in HTS is being investigated [52], which would increase throughput and reduce reagent cost. Statistical experimental design tools are being explored to improve the ro-

bustness of assays [53]. New recombinant microorganisms are being studied to screen for non-antibiotic compounds [54]. A sensitive colorimetric assay for *in vitro* molecular recognition using polymeric artificial membranes has been described [56,57,58]. These membranes, which contain a ligand, can be polymerized into liposomes. These liposomes change their chromatic properties on binding to a solubilized target such as a receptor. Developments in scanning probe microscopy for screening and drug development [59,60] are quite exciting because the molecular interaction could be detected without labeling the target or the ligand.

New analytical devices are also being developed. A detection device based on an amperometric sensor chip [62] and an amperometric electrode probe [63] has been described. The microarray technology that has been developed for analyzing gene expression [65], and other analytical methods used in characterizing combinatorial libraries [66-69], could be adapted for medium-throughput screening applications.

The science of HTS is undergoing explosive growth due to rapid developments in assay technology. Major trends include the development of nonisotopic detection systems and the use of cell-based assays. Miniaturization of assay technologies coupled with automation of high-throughput combinatorial synthesis is helping to set the stage for screening 50-100,000 samples/day in an ultra-HTS mode. Bioinformatics systems to collect, analyze, manipulate and store the massive amount of data are also being rapidly developed. When these capabilities are realized, the multitude of targets derived from the human genome effort can be screened, using large numbers of structurally diverse libraries to generate selective and potent lead compounds. It is also anticipated that the technologies developed will greatly contribute to efficient design of secondary and tertiary assays used to determine structure-activity relationships. The net effect would be the ready availability of multiple, high quality leads to develop novel therapies for the treatment and prevention of disease.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Oliver SG: From DNA sequence to biological function. *Nature* 1996, 379:597-600.
2. Baum R: Combinatorial chemistry. *Chem Eng News*, February 12, 1996:28-54.
3. Broach JR, Thorne J: High throughput screening for drug discovery. *Nature* 1996, 384:14-16.
- This is an excellent overview of high-throughput screening (HTS) with its advantages and disadvantages in early drug discovery. A good account of problems encountered with *in vitro* and cell-based assays are also given. Future developments in HTS are concisely documented.
4. Janzen WP: High throughput screening as a discovery tool in the pharmaceutical industry. *Lab Robotics Automation* 1996, 8:261-265.
- An industrial perspective on steps leading up to the implementation of high-throughput screening is given in this paper. Topics covered include organizational interactions, the screening paradigm employed, and how assays are converted to screens.
5. Fernandes PB: Letter from the society president. *J Biomol Screening* 1997, 2:1.
6. Burbaum JJ, Sigal NH: New technologies for high-throughput screening. *Curr Opin Chem Biol* 1997, 1:72-78.
- An excellent review of new technologies for high-throughput screening. The paper contains references on nonradioactive assay technologies, screening methods for combinatorial libraries and issues associated with assay miniaturization.
7. Sittampalam GS, Iversen PW, Boadt JA, Kahl SD, Bright S, Zock JM, Janzen WP, Lister MD: Design of signal windows in high throughput screening assays for drug discovery. *J Biomol Screening* 1997, 2:169-169.
- This paper describes the concept of signal windows, which provides a degree of separation between measured signals. The size of the signal window is a critical performance parameter (in high-throughput screens) which impacts the identification of active compounds ('hits') in the presence variability.
8. Hook D: Ultra high throughput screening - a journey into Nanoland with Gulliver and Alice. *Drug Discov Tech* 1996, 1:267-268.
9. Braunwalder AF, Wennogle L, Gay B, Lipson KE, Sills MA: Application of scintillation microtiter plates to measure phosphopeptide interactions with GRB2-SH2 binding domain. *J Biomol Screening* 1996, 1:23-28.
10. Cole JL: Approaches to high volume screening assays of viral polymerases and related proteins. *Methods Enzymol* 1996, 275:310-328.
11. Cook ND: Scintillation proximity assay: a versatile high throughput screening technology. *Drug Discov Tech* 1996, 1:287-294.
- A good account of the basic principles of the scintillation proximity assay and its specific applications in high-throughput screening, with 70 references cited.
12. Kahl SD, Hubbard FR, Sittampalam GS, Zock JM: Validation of a high throughput scintillation proximity assay for 5-hydroxytryptamine_{1g} receptor binding activity. *J Biomol Screening* 1997, 2:33-40.
- This paper describes the development of a scintillation proximity assay for receptor binding studies in high-throughput mode. Validation concepts and factors affecting robustness of the assays in high-throughput screening mode are addressed in detail.
13. Lerner CG, Chiang Saiki AY, Mackinnon CA, Xuei X: High throughput screen for inhibitors of bacterial DNA topoisomerase I using scintillation proximity assay. *J Biomol Screening* 1996, 1:135-143.
14. Baker CA, Poorman RA, Kozdy FI, Staples DJ, Smith CW, Elhammer AP: A scintillation proximity assay for UDP-GalNAc:polypeptide, N-Acetylgalactosaminyltransferase. *Anal Biochem* 1996, 238:20-24.
15. Baum EZ, Johnston SH, Bernheim GA, Gluzman Y: Development of scintillation proximity assay for human cytomegalovirus protease using ³²P-phosphorus. *Anal Biochem* 1996, 237:129-134.
- This paper demonstrates the use of ³²P as a label in a scintillation proximity assay system. The authors demonstrate the utility of a simple 'mix and measure' assay to screen for protease inhibitors.
16. Sullivan E, Hensley P, Pickard A: Development of a scintillation proximity assay for calcineurin phosphatase activity. *J Biomol Screening* 1997, 2:19-23.
17. De Serres M, McNulty MJ, Christensen L, Zon G, Findlay JWA: Development of a novel scintillation proximity competitive hybridization assay for the determination of phosphorothioate antisense oligonucleotide plasma concentrations in a toxicokinetic study. *Anal Biochem* 1996, 233:228-233.
18. Sonatore LM, Wisniewski D, Frank LJ, Cameron PM, Hermes JD, Marcy AI, Selowe SP: The utility of FK506-binding protein as a fusion partner in scintillation proximity assays: application to SH2 domains. *Anal Biochem* 1996, 240:286-297.

19. Chan T, Repetto B, Chizzonite R, Pullar C, Burghardt C, Dharm E, Zhao Z, Carroll R, Nunes P, Basu M et al: Interaction of phosphorylated Fc γ R2b immunoglobulin receptor tyrosine activation motif-based peptides with dual and single SH2 domains of p72^{src}. *J Biol Chem* 1998, 271:25308-25316.
 20. Patel S, Harris A, O'Beirne G, Cook ND, Taylor CW: Kinetic Analysis of Inositol triphosphate binding to pure inositol triphosphate receptors using scintillation proximity assay. *Biochem Biophys Res Commun* 1998, 221:821-825.
 21. Fox S: heralding a new era of cell-based assays. *Pharm Forum* 1998, 6:1-3.
 22. Harris DW, Kenrick MK, Pither RJ, Anson JG, Jones DA: Development of a high volume *in situ* mRNA hybridization assay for the quantification of gene expression utilizing scintillating microplates. *Anal Biochem* 1998, 243:249-258.
- A unique application of scintillation proximity assay using scintillating microplates to quantify mRNA *in situ*. This method detects mRNA transcripts at the level of 10-20 copies/cell, and is 20-fold more sensitive than Northern blotting. The authors demonstrate the utility of a high throughput approach to quantify gene expression.
23. Brown MP, Royer C: Fluorescence spectroscopy as a tool to investigate protein interactions. *Curr Opin Biotechnol* 1997, 8:45-49.
 24. Glazer AN: Recent advances in fluorescence labeling, detection and visualization. *BioRadiations* 1997, 98:4-8.
- A good review of the fundamentals of fluorescence techniques for biological measurements. Detection techniques described are easily amenable to high-throughput screening assays. A section on recent advances describe developments in fluorescence resonance energy transfer (FRET) and fluorescence *in situ* hybridization (FISH) technologies, membrane potential measurements and karyotyping human chromosomes. Over 40 references are cited on various applications.
25. Czamk AW: Desperately seeking sensors. *Chem Biol* 1995, 2:423-428.
 26. Wang GT, Matsuyoshi E, Huffaker JE, Kraft GA: Design and synthesis of new fluorogenic HIV protease substrates based on resonance energy transfer. *Tetrahedron Lett* 1991, 31:6493-6496.
 27. Mathis G: Probing molecular interactions with homogeneous techniques based on rare earth cryptates and fluorescence energy transfer. *Chin Chem* 1995, 41:1391-1397.
 28. Kolb JM, Yamanaka G, Manly SP: Use of a novel homogeneous fluorescent technology in high throughput screening. *J Biomol Screening* 1998, 1:203-210.
 29. Gonzalez JE, Tsien RY: Voltage sensing by fluorescence resonance energy transfer. *Biophys J* 1995, 69:1272-1280.
 30. Schroeder KS, Neagle BD: FLIPR: A new instrument for accurate, high throughput optical screening. *J Biomol Screening* 1996, 1:75-80.
- This paper describes an instrument that can simultaneously read fluorescence signals from cells in all 96 wells of a microtiter plate. Kinetic updates can be obtained in less than one second in all wells, and the equipment allows for measurement of transient fluxes in intracellular Ca²⁺, pH and membrane potential.
31. Waggoner A, Taylor L, Seadler A, Dunlay T: Multiparameter fluorescence imaging microscopy: reagents and instruments. *Hum Pathol* 1998, 27:494-502.
 32. Jameson DM, Sawyer WH: Fluorescence anisotropy applied to biomolecular interactions. *Methods Enzymol* 1995, 248:283-300.
 33. Lundblad JR, Laurance M, Goodman RH: Fluorescence polarization analysis of protein-protein interactions. *Mol Endocrinol* 1998, 10:607-612.
 34. Checovich WJ, Bolger RE, Burke T: Fluorescence polarization - a new tool for cell and molecular biology. *Nature* 1995, 375:254-258.
 35. Levine LM, Michener ML, Toth MV, Holwerda BC: Measurement of specific protease activity utilizing fluorescence polarization. *Anal Biochem* 1997, 247:83-88.
- This is the first report describing the use of fluorescence polarization in a high-throughput mode using a modified 96-well plate reader. The peptide substrate for the human cytomegalovirus protease was labeled with biotin and DTAF (5,6-dichlorotriazin-2-yl)amino)fluorescein). The uncleaved substrate, when bound to avidin, produced a high polarization value; hence, the presence of inhibitors in the mixture can be easily identified.
36. Jolley ME: Fluorescence polarization assays for the detection of proteases and their inhibitors. *J Biomol Screening* 1998, 1:33-38.
 37. Schade SZ, Jolley ME, Sarauer BJ, Simonson LG: BODIPY- α -Casein, a pH-independent protein substrate for protease assays using fluorescence polarization. *Anal Biochem* 1998, 243:1-7.
 38. Lynch BA, Loiacono LA, Tong CL, Adams SE, MacNeil IA: A fluorescence polarization based Src-SH2 binding assay. *Anal Biochem* 1997, 247:77-82.
- Application of fluorescence polarization in a high-throughput mode to a protein-peptide interaction assay involving Src-SH2 domain. Carboxyfluorescein without a linker was used as the label for the peptide probe to minimize propeller effect. The assay tolerated up to 20% dimethyl sulfoxide (DMSO), a common solvent used to dissolve compounds tested in high-throughput screening.
39. Sterner S, Henco K: Fluorescence correlation spectroscopy (FCS) - A highly sensitive method to analyze drug/target interactions. *J Recept Signal Transduct Res* 1997, 17:511-520.
- This paper describes the use of fluorescence correlation spectroscopy to measure molecular interactions in a homogeneous mode. It has been developed to accommodate measurements in 96- and 384-well microtiter plates, and is a good candidate for high-throughput screening applications.
40. Rigler R: Fluorescence correlations, single molecule detection and large number screening. Applications in biotechnology. *J Biotechnol* 1995, 41:177-186.
 41. Rauer B, Neumann E, Widgren J, Rigler R: Fluorescence correlation spectroscopy of the interaction kinetics of tetramethylrhodamine α -bungarotoxin with *Torpedo californica* acetylcholine receptor. *Biophys Chem* 1998, 58:3-12.
 42. Sanzubi E, Yanofsky SD, Barrett RW, Denaro M: A cell-free, nonisotopic, high-throughput assay for inhibitors of type-1 interleukin-1 receptor. *Anal Biochem* 1998, 237:70-75.
 43. Ross P, Gorman J, Kurtz S, Patel P, Fernandes P: The successful partnership of biotechnology based screen development with high throughput screening. *Network Science* 1998, 2(Sept):1-12. On the World Wide Web URL: <http://www.ewod.com/netsci/science/Screening/featureeb/html>
- A comprehensive review of cell-based assay systems available for high-throughput screening applications. Targets reviewed include ion channels, G-protein-coupled receptors, tyrosine kinase receptors, intracellular receptors, protein-protein interactions and proteases. Over 40 references are cited.
44. Dhundale A, Goddard C: Reporter assays in high throughput screening laboratory: A rapid and robust first look? *J Biomol Screening* 1998, 1:115-118.
 45. Suto CM, Ignar DM: Selection of an optimal reporter gene for cell-based high throughput screening assays. *J Biomol Screening* 1997, 2:7-9.
 46. Bronstein I, Fortin J, Stanley PE, Stewart GSAB, Kricka LJ: Chemiluminescent and bioluminescent reporter gene assays. *Anal Biochem* 1994, 219:169-181.
 47. Hastings WJ: Chemistries and colors of bioluminescent reactions: a review. *Gene* 1998, 173:5-11.
 48. Lehel C, Daniel-Haskani S, Brassau M, Sindrovi B: A chemiluminescent microplate assay for sensitive detection of protein kinase activity. *Anal Biochem* 1997, 244:340-348.
 49. Kolb AJ, Neumann K: Luciferase measurements in high throughput screening. *J Biomol Screening* 1998, 1:85-88.
 50. Bran MR, Messier T, Doman C, Lannigan D: Cell-based assays for G-protein-coupled/Tyrosine kinase coupled receptors. *J Biomol Screening* 1998, 1:43-45.
 51. Rizzuto R, Brini M, De Giorgi F, Rossi R, Heim R, Tsien RY, Pozzan T: Double labelling subcellular structures with organelle-targeted GFP mutants *in vivo*. *Curr Biol* 1998, 8:183-188.
 52. Janzen B, Domanico P: The 384-well plate: pros and cons. *J Biomol Screening* 1998, 1:83-84.
 53. Lutz MW, Menius A, Choi TD, Laskody RG, Domanico PL, Goetz AS, Saussy DL: Experimental design for high-throughput screening. *Drug Discov Tech* 1998, 1:277-288.
 54. Klein RD, Geary GG: Recombinant microorganisms as tools for high throughput screening for non antibiotic compounds. *J Biomol Screening* 1997, 2:41-49.

55. Webb SA, Hurakainen P: Transcription-specific assay for quantifying mRNA: A potential replacement for reporter gene assays. *J Biomol Screening* 1998, 1:119-121.
56. Charych DH, Nagy JO, Spevak W, Bednarski MD: Direct colorimetric detection of receptor-ligand interaction by a polymerized bilayer assembly. *Science* 1993, 261:585-588.
57. Charych D, Cheng Q, Reichart A, Kuziemko G, Stroh M, Nagy JO, Spevak W, Stevens RC: A 'litmus test' for molecular recognition using artificial membranes. *Chem Biol* 1998, 3:113-120.
- A unique, membrane-based colorimetric system that detects molecular interactions is described. Gangliosides that specifically bind cholera toxin, *Escherichia coli* enterotoxin and botulinum neurotoxin were incorporated into a polydiacetylene membrane. The polymerized membrane containing gangliosides is blue, and turns red when the toxin is added. The response is sensitive, specific and selective. An excellent technology for high-throughput screening applications.
58. Spevak W, Foxall C, Charych DH, Dasgupta F, Nagy JO: Carbohydrates in an acidic multivalent assembly: nanomolar P-selectin inhibitors. *J Med Chem* 1996, 39:1018-1020.
59. Allen S, Davies MC, Roberts CJ, Tendler SJB, Williams PM: Atomic force microscopy in analytical biotechnology. *Trends Biotechnol* 1997, 15:101-105.
60. Troy CT, Abrams SB: Scanning force microscopy helps in the design of cancer drugs. *Biophoton Int* 1998, Sept/Oct:52-53.
61. Paborsky LR, Dunn KE, Gibbs CS, Dougherty JP: A nickel chelate microtiter plate assay for six histidine-containing proteins. *Anal Biochem* 1998, 234:60-65.
62. Weiss-Wichert CH, Smetarzo M, Valina-saba M, Schalkhammer TH: A new analytical device based on gated ion channels: A peptide channel biosensor. *J Biomol Screening* 1997, 2:11-18.
63. Brecht A, Burckhardt R, Ricket J, Stemmler I, Schuetz A, Fischer S, Friedrich T, Gaugitz G, Goepel W: Transducer-based approaches for parallel binding assays in HTS. *J Biomol Screening* 1998, 1:191-201.
64. Tyagi S, Kramer FR: Molecular beacons: probes that fluoresce upon hybridization. *Nat Biotechnol* 1996, 14:303-308.
65. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW: Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997, 94:2160-2165.
66. Nicolaou KC, Xiao XY, Parandoosh Z, Senyeli A, Nova MP: Radiofrequency encoded combinatorial chemistry. *Angew Chem Int Ed* 1995, 34:2289-2291.
67. Fitzgerald MC, Harris K, Shevlin CG, Szardak G: Direct characterization of solid phase resin-bound molecules by mass spectrometry. *Bioorg Med Chem Lett* 1998, 6:979-982.
68. Chu YH, Dunnayevskiy YM, Kirby DP, Vouros P, Karger BL: Affinity capillary electrophoresis-mass spectrometry for screening combinatorial libraries. *J Am Chem Soc* 1998, 118:7827-7835.
69. Evans DM, Williams KP, McGuinness B, Tarr G, Regnier F, Afeyan N, Jindal S: Affinity-based screening of combinatorial libraries using automated, serial-column chromatography. *Nat Biotechnol* 1998, 14:504-507.

Exhibit 34

This paper was presented at the National Academy of Sciences colloquium "Proteolytic Processing and Physiological Regulation," held February 20–21, 1999, at the Arnold and Mabel Beckman Center in Irvine, CA.

The structure of the human β II-tryptase tetramer: Fo(u)r better or worse

CHRISTIAN P. SOMMERHOFF*[†], WOLFRAM BODE[‡], PEDRO J. B. PEREIRA[‡], MILTON T. STUBBS[§],
JÖRG STÜRZEBECKER[¶], GERD P. PIECHOTTKA*, GABRIELE MATSCHINER*, AND ANDREAS BERGNER[‡]

*Abteilung Klinische Chemie und Klinische Biochemie in der Chirurgischen Klinik und Poliklinik, Klinikum Innenstadt der Ludwig-Maximilians-Universität, Nußbaumstrasse 20, D-80336 Munich, Germany; [‡]Abteilung für Strukturforschung, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany; [§]Institut für Pharmazeutische Chemie der Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany; and [¶]Klinikum der Universität Jena, Zentrum für Vaskuläre Biologie und Medizin, Nordhäuserstrasse 78, D-99089 Erfurt, Germany

ABSTRACT Trypsases, the predominant serine proteinases of human mast cells, have recently been implicated as mediators in the pathogenesis of allergic and inflammatory conditions, most notably asthma. Their distinguishing features, their activity as a heparin-stabilized tetramer and resistance to most proteinaceous inhibitors, are perfectly explained by the 3-Å crystal structure of human β II-tryptase in complex with 4-amidinophenylpyruvic acid. The tetramer consists of four quasiequivalent monomers arranged in a flat frame-like structure. The active centers are directed toward a central pore whose narrow openings of approximately 40 Å × 15 Å govern the interaction with macromolecular substrates and inhibitors. The tryptase monomer exhibits the overall fold of trypsin-like serine proteinases but differs considerably in the conformation of six surface loops arranged around the active site. These loops border and shape the active site cleft to a large extent and form all contacts with neighboring monomers via two distinct interfaces. The smaller of these interfaces, which is exclusively hydrophobic, can be stabilized by the binding of heparin chains to elongated patches of positively charged residues on adjacent monomers or, alternatively, by high salt concentrations *in vitro*. On tetramer dissociation, the monomers are likely to undergo transformation into a zymogen-like conformation that is favored and stabilized by intramonomer interactions. The structure thus provides an improved understanding of the unique properties of the biologically active tryptase tetramer in solution and will be an incentive for the rational design of mono- and multi-functional tryptase inhibitors.

Human mast cell tryptases (EC 3.4.21.59) comprise a family of trypsin-like serine proteinases closely related in sequence that are derived from ≥ 3 nonallelic genes (1, 2). Tryptases (at least isoenzymes α I, β I, β II, and β III) are highly and selectively expressed in mast cells and to a lesser extent in basophils (3, 4). Only β -tryptases, however, appear to be activated intracellularly and stored in secretory granules (5, 6), accumulating to much larger amounts than any other of the granule-associated serine proteinases of leukocytes and lymphocytes. On mast cell activation, β -tryptases are secreted bound to heparin in diverse allergic and inflammatory conditions ranging from asthma and rhinitis to psoriasis and multiple sclerosis. Various studies performed in animals and humans have provided considerable evidence that tryptases are directly involved in the pathogenesis of asthma (7–9), a hypothesis also supported by apparent genetic links of tryptases to airway reactivity (10, 11).

Several unique properties distinguish tryptases from other trypsin-like proteinases (reviewed in refs. 12 and 13). Most notably, tryptases are enzymatically active in the form of a noncovalently linked tetramer. The tetramer is stabilized by association with negatively charged aminoglycans such as heparin or high ionic strength conditions *in vitro*. On dissociation, reversible only under certain conditions, the monomers lose activity, apparently because of transition into a zymogen-like state (14, 15). This mechanism is thought to govern tryptase activity *in vivo*. With the exception of the "atypical" Kazal-type inhibitor leech-derived tryptase inhibitor (LDTI) (16, 17), human tryptases are resistant to inhibition by proteinaceous inhibitors. In accordance with their trypsin-like activity, tryptases efficiently hydrolyze a number of peptide substrates including the neuropeptides "vasoactive intestinal peptide" and "peptide histidine methionine" (18). Few macromolecular substrates are cleaved, however, leading to the activation of prostromelysin, prourokinase, and the proteinase-activated receptor-2 (19–21) and the inactivation of fibronectin and of the procoagulant functions of high molecular-mass kininogen and fibrinogen (22–24).

These distinguishing features are well explained by the crystal structure of the human lung β II-tryptase tetramer, whose overall architecture has been summarized recently (25). Here, we describe the identification of the tetramer within the crystal packing, the detailed structure of the monomers, and their interactions in the tetramer. In addition, structural features likely to favor a zymogen-like conformation of isolated monomers and models of the interaction with stabilizing heparin proteoglycans and inhibitors are presented.

Identification of the Relevant Tryptase Tetramer. In the x - y plane of the tryptase crystals, the tryptase monomers are arranged in flat rectangular tetrameric aggregates that form extended protein layers (Fig. 1a). Within these layers, each tetramer is rotated about the crystallographic a - and b -axes by $\approx 7^\circ$, in agreement with the self-rotation function. The tetramers appear well separated from their neighbors in one direction (x -direction in Fig. 1a) but are in somewhat closer contact in the perpendicular direction (y in Fig. 1a). In the z -direction, the tetramers are stacked along the crystallographic 4_1 screw axis. Because of the 7° tilt of each tetramer from the x - y plane, their projections (Fig. 1b) alternate between leaning to the left, being horizontal, and leaning to the right, respectively, giving rise to a 7° precession motion of the

Abbreviations: APPA, 4-amidinophenylpyruvic acid; LDTI, leech-derived tryptase inhibitor.

Data deposition: The atomic coordinates have been deposited in the Protein Data Bank, www.rcsb.org (PDB ID code 1A0L).

[†]To whom reprint requests should be addressed. E-mail: sommerhoff@clinbio.med.uni-muenchen.de.

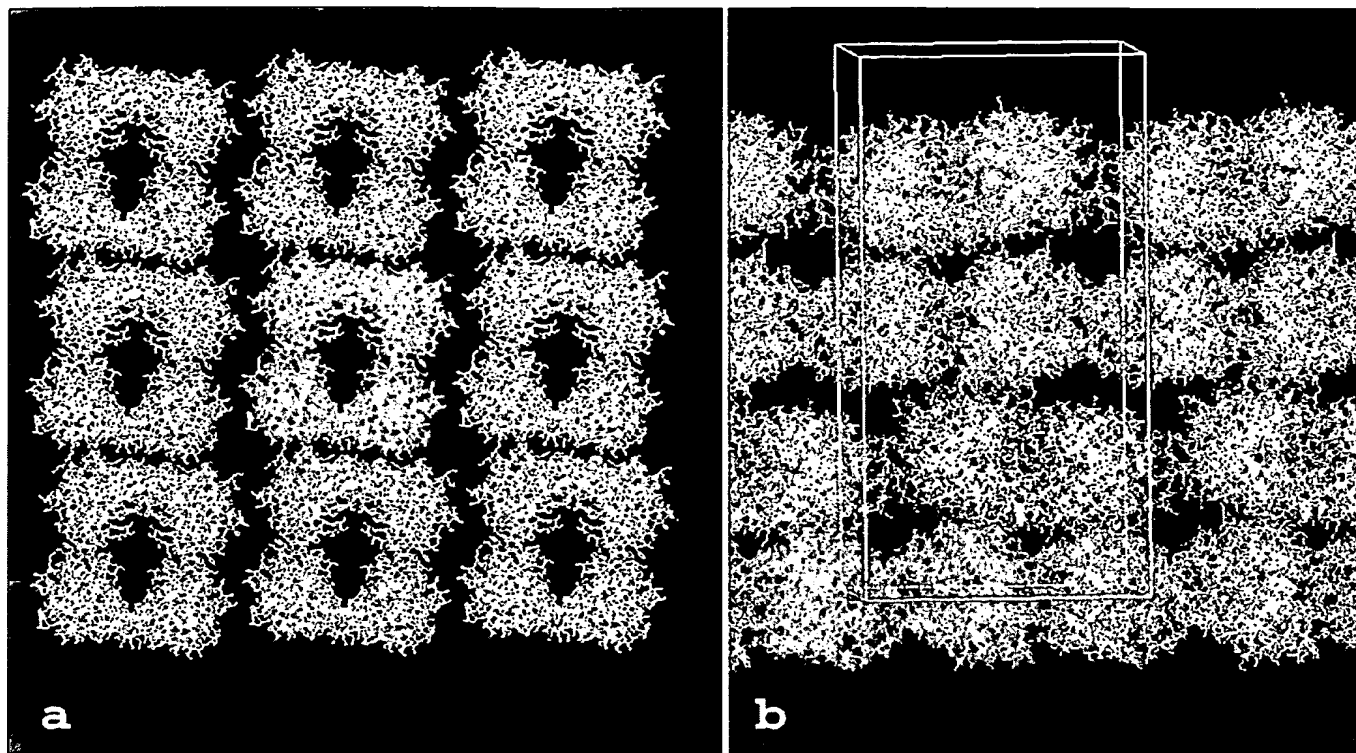


FIG. 1. Packing of the human BII tryptase crystal. (a) View along the z -axis showing one layer of tryptase molecules in the x - y plane. The tryptase monomers are grouped into tetrameric aggregates that form extended sheets. Each of these tryptase tetramers is clearly delimited from its neighbors in both directions. A "reference" tetramer is shown in red for simplicity. (b) View across the z -axis. In the z direction, layers of tetramers are stacked on each other via much more usual crystal contacts. The local 2-fold symmetry axis is tilted from the z direction by $\approx 7^\circ$, causing increased crystal-stabilizing contacts between layers stacked in the z -direction. One unit cell ($82.9 \times 82.9 \times 172.9 \text{ \AA}$), occupied by four tryptase tetramers, is indicated by a white bordered box.

local (2-fold; see below) rotation axis along the crystallographic 4_1 screw axis. The largely complementary interaction surfaces between the monomers of the tetramer are typical for intersubunit contacts, whereas neighboring tetramers interact with one another via much more usual crystal contacts. Thus, within a tetramer, monomer A (Fig. 2) interacts with monomers B and D via interfaces of sizes 540 \AA^2 and $1,075 \text{ \AA}^2$, respectively (solvent inaccessible surface probed by using a sphere of 1.4-\AA radius; Collaborative Computational Project No. 4 suite). In contrast, the four monomers of one given tetramer interact with monomers from neighboring tetramers via interfaces of less than 280 \AA^2 (in the x - y plane) and 265 \AA^2 (along the z -axis), respectively. The contacts between tetramers include a number of hydrogen bonds and six unique salt bridges and thus are qualitatively similar to those usually observed in typical crystal contacts.

These packing considerations suggest that the tetramer emphasized in Fig. 1 represents the enzymatically active tetramer of human β -tryptase. This tetramer selection is supported by the finding that the six loops that deviate most from the structures of other trypsin-like proteinases are all involved in forming monomer-monomer contacts within a tetramer. More important, this unique tetramer perfectly explains the distinguishing properties of tryptase in solution, e.g., the resistance to proteinaceous inhibitors other than LDTI, the unusual substrate specificity, and the stabilization by the binding of heparin-like glycosaminoglycans (see below).

Overall Tetramer Structure. In the tryptase tetramer, monomers (arbitrarily assigned A, B, C, and D in Fig. 2) are positioned at the corners of a flat rectangular frame leaving a continuous central pore. The tetramer displays almost perfect 222 symmetry that, however, is not exact because of the crystallographically asymmetric environment and an imperfect

internal packing (see below). The horizontal and the vertical 2-fold axes, which cross each other in the center of the tetramer, relate monomers A to B and C to D, or A to D and B to C, respectively. The third 2-fold symmetry axis relating monomers A to C and B to D is arranged virtually perpendicular to the other 2-fold axes and runs almost through their point of intersection in the central pore.

The active centers of the four monomers are directed toward the central pore (Fig. 2). This pore exhibits a rectangular cross section and is twisted by $\approx 30^\circ$ about the tetramer axis. It possesses two narrow openings of dimension $40 \text{ \AA} \times 15 \text{ \AA}$, and widens in its central part to a cross section of $50 \text{ \AA} \times 25 \text{ \AA}$, just large enough for elongated peptides of the diameter of an α -helix to thread through the exits and to interact with the active sites. Both pore entrances are partially obscured by the 147-loops (see below), which project from each of the monomers but on alternative entrance sides, so that only two diagonally arranged active centers can be viewed directly (Fig. 2). With 33 basic (including 12 His residues) and 24 acidic residues per monomer, human tryptase exhibits an average percentage of charged residues comparable to related serine proteinases, but is only slightly positively charged at neutral pH. These charges are not evenly distributed along the molecular surface, however. Rather, negatively charged residues cluster preferentially on the inner pore-facing surface, conferring the pore with a quite negative electrostatic potential, and along the peripheral A-D (and B-C) edges. In contrast, the A-B (and C-D) peripheries and one front side of the monomer surface are positively charged and probably are involved in heparin binding (see below and Fig. 6).

Monomer Structure. The tryptase monomer exhibits the typical β -strand-dominated fold seen in other trypsin-like serine proteinases. The core is made by two six-stranded

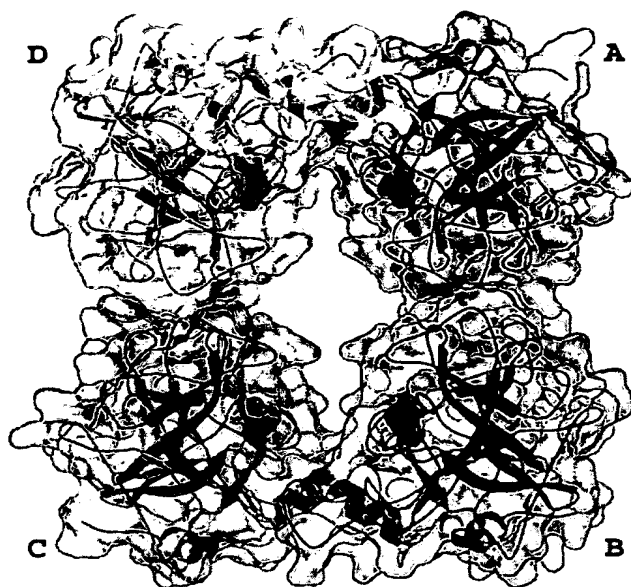


FIG. 2. Overall structure of the tryptase tetramer. The four monomers A, B, C, and D (clockwise) are shown as blue, red, green, and yellow ribbons, each surrounded by a semitransparent surface. The inhibitor molecules APPA are given as orange CPK models, each binding into one of the four S1 specificity pockets.

β -barrels that are packed together and further clamped by three transdomain segments (Fig. 3). This core structure is covered by a number of polypeptide loops, a short α -helical turn (Ala-55–Gly-66, not shown in Fig. 3*a*), and two regular α -helices, the so-called “intermediate helix” (Glu-164–Leu-173A) and the C-terminal helix (Arg-230–Val-242). The catalytic residues Ser-195, His-57, and Asp-102 (chymotrypsinogen numbering) are located in the junction between both barrels. The active-site cleft runs perpendicular to this barrel junction. In the “standard orientation” shown in Fig. 3, this cleft runs approximately horizontally across the molecular surface facing the viewer and is ready to accommodate and bind extended peptide substrates extending from left to right. One hundred sixty-two and 168 residues of the tryptase monomer are topologically equivalent to the archetypal proteinases chymotrypsin (26) and trypsin (27), respectively, with an rms deviation of their α -carbon atoms of 0.65 Å for both comparisons. The numbering of the tryptase residues given in this article is predominantly based on the equivalence with chymotrypsinogen (28) and at only a few trypsin-characteristic sites on that with trypsin (27).

In detail, however, the topology of the tryptase monomers deviates significantly from these reference proteinases (Fig. 3*b*), probably more than any other trypsin-like serine proteinase. In particular, six surface loops that border and shape the active-site cleft are unique (Fig. 3*a*). These loops comprise the 147-loop (including the 152-“spur”), the 70- to 80-loop, the 37-loop, the 60-loop, the 97-loop, and the 173-loop (Fig. 3*a*). The 147-loop, which together with Gln-192 forms the rather acidic southern wall of the active-site cleft, is shortened by one residue in its initial part, but contains a two-residue insertion (Pro-152–Pro-152A–cisPro-152B–Phe-153–Pro-154) in its proline-rich and hydrophobic 152-spur. The neighboring 70- to 80-loop to the east, which in the calcium-binding serine proteinases winds around a stabilizing calcium ion (27), is three residues shorter and more compact in tryptase. It is probably not designed for calcium binding, in spite of topologically similar liganding groups; Glu-70 and Asp-80, involved in a partially buried salt bridge cluster with Arg-34, are

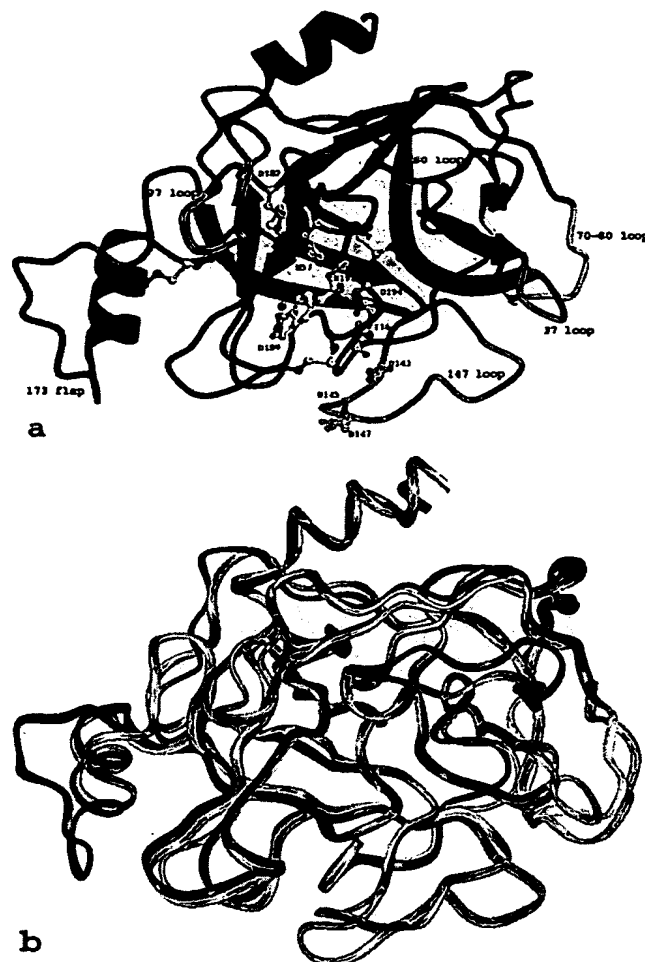


FIG. 3. The tryptase monomer in standard orientation, i.e., as seen approximately from the middle of the central pore of the tetramer toward the active site of monomer A (represented by Ser-195, His-57, and Asp-102). (a) Ribbon representation of a tryptase monomer. The amidino group of the APPA molecule interacts with Asp-189 in the S1 pocket. Ser-195 O- γ is bound covalently to the APPA carbonyl group forming a hemiketal. The six unique surface loops of tryptase that surround the active site and are engaged in intermonomer contacts are shown in special colors, namely (anticlockwise) the 147-loop (light blue), the 70- to 80-loop (yellow), the 37-loop (orange), the 60-loop (magenta), the 97-loop (green), and the 173-loop (red). All other tryptase segments are given in dark blue. The side chains of the catalytic triad residues as well as Asp-143, Asp-145, and Asp-147 in the acidic 147-loop are shown as a ball-and-stick model. (b) Overlay of the structures of the tryptase monomer and bovine trypsin, both given as ropes. The color-coding of tryptase is as in *a*, whereas trypsin is shown in gray. The most relevant deviations from the trypsin backbone appear in the colored loop regions of tryptase.

oppositely arranged to the two calcium-binding Glu residues in trypsin. The 37-loop, above the 70- to 80-loop, possesses two additional residues (Pro-37A and Tyr-37B), which bulge away from the loop axis. The adjacent 60-loop, with five inserted residues, turns away from the cleft abruptly to the north, where it kinks at cisPro-60A to approach the general main chain course of other serine proteinases. At position 69, a buried Arg replaces the Gly residue that is strictly conserved in most other homologous proteinases, allowing for a special conformation. Although the 97-loop, at the northern rim of the cleft, contains the same number of residues as other serine proteinases, it differs considerably in conformation. The N-terminal part is shortened by two residues between positions 96 and 97, thus placing Ala-97 in the position normally occupied by residue 99,

whereas its C-terminal part makes an unusual extra helical turn before arriving at Asp-102. By far the largest insertion, with nine residues, occurs in the 173-loop. After the unusually long three-turn intermediate helix, the 10 residues from His-173 to Val-1731 form an exposed flap centered around the imidazole side chain of His-173.

With 245 amino acid residues, the trypsin monomer possesses 15 and 22 residues more than the B-chains of chymotrypsin and trypsin, respectively. Compared with chymotrypsinogen, most of these extra residues present in all trypsinases known so far are inserted in the 37-loop (two residues), the 60-loop (+5), the 147-loop (+1), the 173-loop (+9), at position 221A (+1) and at the C terminus (+1), whereas the 70- to 80-loop (-3) and the 214- to 220-loop (-1, as in all trypsin-like serine proteinases) are shorter. On the reverse side, the largely hydrophobic cluster of four Trp residues (Trp-27, -29, -207, and -137) is noteworthy. Only the indole moieties of the latter two Trp are significantly exposed to the surface. At the C terminus, only the main chain atoms of the two penultimate residues Lys-244 and Lys-245 are well defined by electron density, while the C-terminal Pro-246 could not be located. The side chain of the single N-linked sugar attachment site in human β II-trypsin, Asn-204, extends away from the molecular surface opposite to the active site. Some residual electron density exists distal to its carboxamide group, which is not large enough to account for a covalently linked sugar residue.

As found in almost all trypsin-like serine proteinases [except, e.g., single-chain tissue type plasminogen activator (29)], the N-terminal Ile-16-Val-17 segment is inserted in the Ile-16 pocket, forming a solvent inaccessible salt bridge between its free Ile-16 α -amino group and the carboxylate group of Asp-194. The formation of this salt bridge after activation cleavage creates a functional substrate recognition site by reorienting the Asp-194 side chain from an external position in the zymogen, where it might hydrogen bond to a surface located His-40...Ser-32 pair forming the so-called "zymogen triad," to an internal position in the active molecule (30, 31). This reorientation restructures the surrounding "activation domain," which in trypsin(ogen) mainly includes the linings of the Ile-16 pocket and the S1 specificity pocket (i.e., segments Ile-16-Gly-19, Tyr-184-Asp-194, Gly-216-Asn-223, and Gly-142-Tyr-151), and the "oxyanion hole" formed by the amide groups of Gly-193 and Ser-195 (28, 32, 33). The single-chain zymogen and the activated monomer are adequately described by a two-state model, in which an inactive conformation is in equilibrium with an active form possessing a structured activation domain (31). The partition between both forms depends on environmental conditions such as the endogenous free Ile-16-Val-17 N-terminal segment (34), free Ile-Val dipeptide (31), ligands in the substrate binding site (30, 36), or other effectors such as fibrin with respect to tissue plasminogen activator or tissue factor in the case of Factor VIIa (29, 37). This conformational partition can be influenced by internal molecular groups that stabilize or destabilize one or the other state. Trypsin possesses the zymogen triad residues His-40 and Ser-32, which would stabilize the zymogen state. In addition, the acidic residues Asp-143, Asp-145, and Asp-147 arranged around the Ile-16 cleft could form a negatively charged anchoring site that could compete with the Ile-16 pocket for the Ile-16 α -amino group, thus destabilizing the structured active state of the trypsin monomer. Furthermore, some of the loops in contact with the activation domain of trypsin, such as the long 173-loop or the 70- to 80-loop, which has been shown to be strongly correlated with the equilibrium state in bovine elastase "subunit III" (38), could influence the structured state. The conformation of the trypsin 173-loop, probably held in place in the tetramer by contacts with monomer D, certainly has an effect on the stability of the integrated monomer. Interestingly, tissue factor, thought to support insertion of the N-terminal Ile-16 α -amino terminus of

activated Factor VIIa B-chain on complex formation (37), likewise binds to the 173-loop at the intermediate helix flank (39).

Interfaces. All monomer-monomer contacts within the tetramer are realized via six loops arranged around the active center. These loops, emphasized by special colors in Figs. 3-5, differ fundamentally in their conformation and partly in size from those of other trypsin-like serine proteinases. Monomers A and B interact with one another through the 147-loop, the 70- to 80-loop, and the 37-loop (Fig. 4d). Each 152-spur slots into a cleft formed by the 37- and the 70- to 80-loop of its own monomer and the 152-spur of the opposing neighbor. At the center of the interface, the side chains of Phe-153 and Tyr-75 from each subunit form an approximate tetrahedron (Fig. 5a). The side chain of Tyr-75 from monomer B (D) would clash with the equivalent A (C) side chain if they were arranged in a symmetrical manner. Instead, the phenolic group of Tyr-75 of monomer A turns in the opposite direction, breaking the 2-fold symmetry (see the partial electron density in Fig. 5a). This A-B (C-D) interface is exclusively hydrophobic, with a remarkable number of Tyr and Pro side chains involved, and lacks any intermonomer hydrogen bonds. Toward the pore, the side chains of the two Arg-150 residues oppose one another. The charges of their guanidyl groups presumably make unfavorable energy contributions to the A-B interaction.

Monomer A interacts with monomer D through the entire northern rim consisting of the 173-flap, the 97-loop, and the 60-loop (Figs. 4a and 5b), again via equivalent loops. Both 97-loops rest with their 95-99 segments on one another (Fig. 4a), with both Ile-99 side chains in direct contact. Further toward both peripheries, segment Pro-60A-Asp-60B and the opposing segment Gly-173B-Tyr-173D run antiparallel to one another, forming two-rung antiparallel ladders between Gly-173B-Tyr-173D and Pro-60A-Val-60C (Fig. 5b). Each Tyr-95 aromatic side chain nestles into the bend of the opposing 173-flap, and each Tyr-173D phenolic side chain slots into a hydrophobic cleft made by the 60-loop and the 97-loop of the opposing monomer. In addition, both monomers are cross-connected by salt bridges between Asp-60B and Arg-224 and

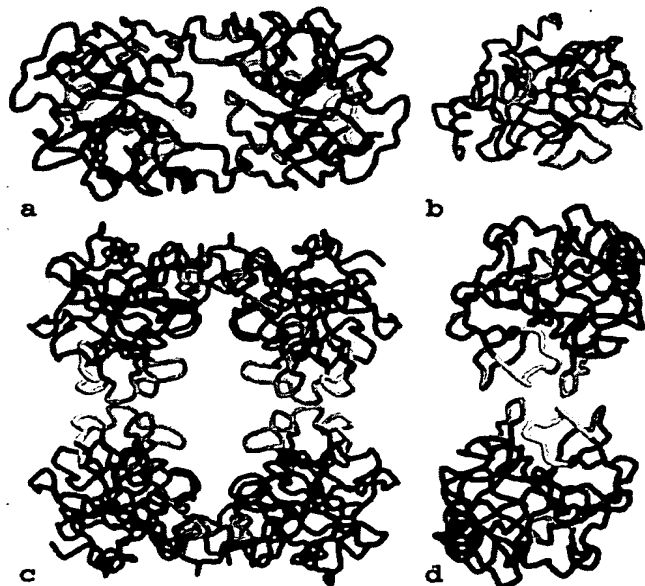


FIG. 4. Loop arrangements in the tetramer. The six special loops engaged in monomer-monomer interactions are shown in the color coding introduced in Fig. 3. (a) The D-A dimer as seen from outside of the tetramer along the local 2-fold axis. (b) The monomer viewed in standard orientation. (c) Front view of the tetramer. (d) The A-B dimer seen from outside of the tetramer along the local 2-fold axis.

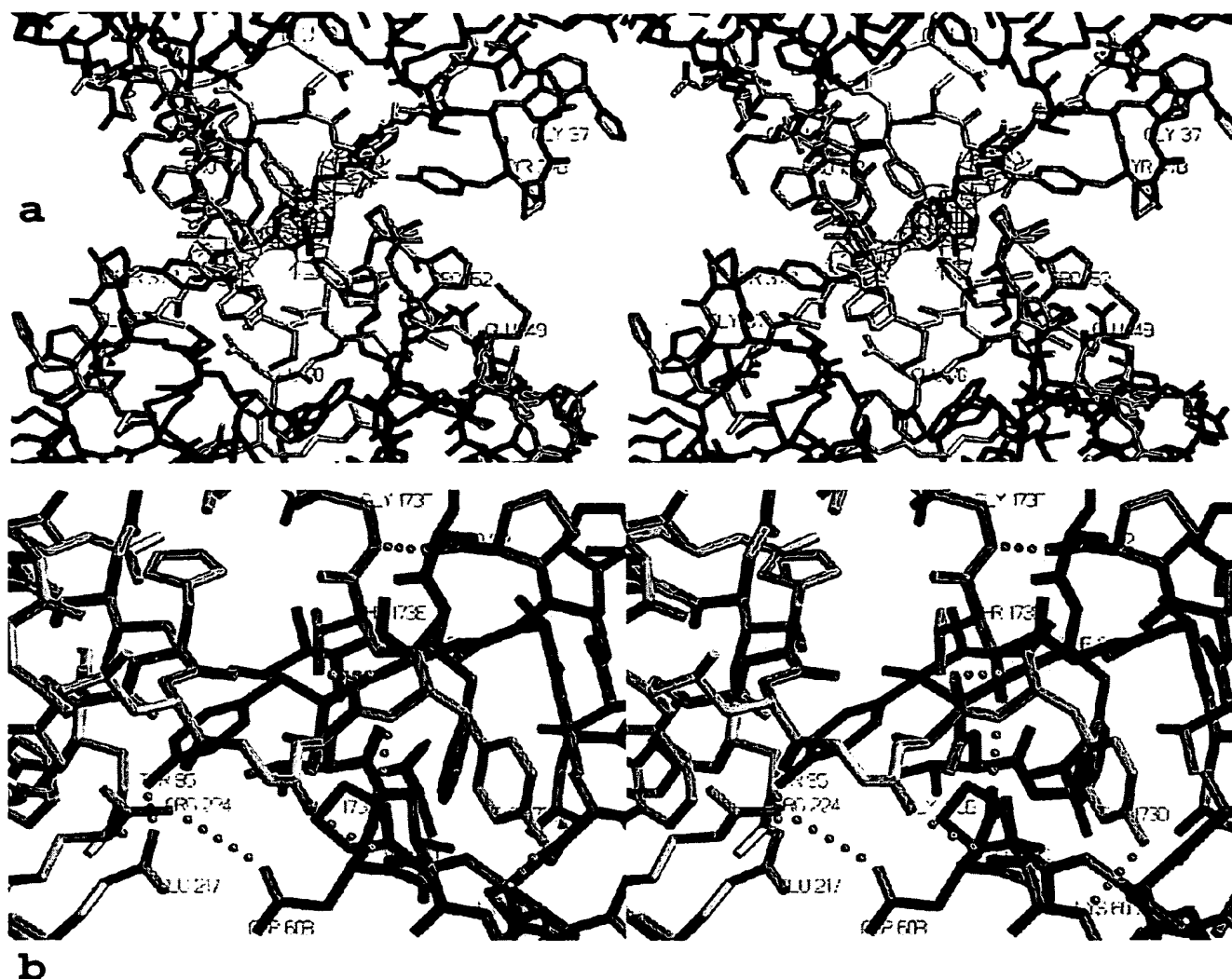


FIG. 5. Stick representation of the contact interfaces between monomers. (a) The AB-interface seen from inside the tetramer along the local 2-fold axis, shown together with the final $2F_o - F_c$ electron density map for both Tyr-75 side chains contoured at 1σ level. The monomers and loops are given in the color coding introduced in Figs. 3 and 4. (b) The AD-interface (half side) observed approximately perpendicular to the local 2-fold axis, shown together with all intermonomer hydrogen bonds and salt bridges (green dots). Segments of monomers A and D are given in blue and yellow, respectively.

by four hydrogen bonds involving both main and side chains (Fig. 5b). Thus, the A-D (and the corresponding B-C) interface comprises a number of polar/charged interactions in addition to several hydrophobic contacts.

The A-B homodimer carries a number of positively charged residues at the periphery, which cluster and form an obliquely oriented two-lobed patch of positive charges that extends toward one of the front sides of each monomer, giving rise to the blue-colored electrostatic potential surfaces in Fig. 6. With an overall length of almost 100 Å, this patch would allow tight electrostatic binding of an extended heparin chain of ≈ 20 sugars running obliquely along the A-B edge as shown in Fig. 6. The length of such heparin chains is in good agreement with the experimentally observed stabilization of the tetramer by heparin fractions of molecular mass 5,500 Da and above (40). On the peripheral surface of the A-D (and the corresponding B-C) homodimer, in contrast, positive charges are counterbalanced by adjacent negative ones.

Interaction with Substrates and Inhibitors. The immediate vicinity of the tryptase active site is quite similar in structure to that of trypsin. The specificity S1 pocket, which opens to the west of the reactive Ser-195 (Fig. 3a), is virtually identical

to that of trypsin and well suited to accommodate P1-Lys and Arg side chains. The 4-amidinophenylpyruvic acid (APPA) molecule inserts into this pocket in the same manner as in the complex with trypsin (41). Thus, its amidino group hydrogen is bonded to both Asp-189 carboxylate oxygens, Gly-219 O and Ser-190 O γ , and its phenyl ring is sandwiched between peptide planes 215–216 and 190–192. Ser-195 O γ bonds to the carbonyl group of the tetrahedral pyruvate part of APPA (Fig. 3a), and hydrogen bonds to His-57 Ne. As indicated by the relatively low equilibrium dissociation constant of the APPA-tryptase complex [K_d 0.71 μ M; (42)], APPA fits well to the tryptase active site. Toward the south of the active site of tryptase, the side chains of Asp-143, Asp-145, and Asp-147 protrude from the relatively flat and hydrophobic southern embankment (Fig. 3a). The resulting negative charge cluster provides a second anchoring point for dibasic synthetic tryptase inhibitors such as bis-benzamidines (17, 42, 43), allowing favorable interactions with a distal basic group such as in pentamidine. The structural basis of the unexpected high affinity of bifunctional inhibitors containing suitably arranged adjacent imidazole moieties such as present in the inhibitor BABIM and closely related analogues (43, 44) has recently been revealed: two nitrogen atoms

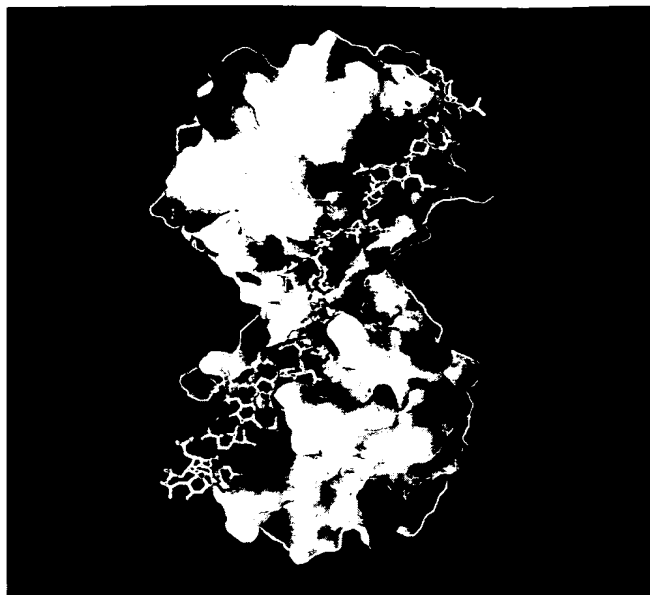


FIG. 6. Model of the binding of a 20-mer heparin-like glycosaminoglycan chain along the A-B edge of the tryptase-tetramer. The solid-surface representation of tryptase indicates positive (blue) and negative (red) electrostatic potential contoured from -4 kT/e to 4 kT/e. The heparin chain (green/yellow/red stick model) is long enough to bind to clusters of positively charged residues on both sides of the monomer-monomer interface, thereby bridging and stabilizing the interface which is exclusively hydrophobic in nature (see Fig. 5a).

of the two methylene-connected benzimidazoles coordinate a zinc ion that also binds to the active-site located Ser-195 O γ and His-57 N ϵ (44). The zinc-mediated binding enhancement of BABIM-like inhibitors is particularly large but not restricted to tryptase.

Toward the east, the substrate-binding site of tryptase is not only bounded by the side chains of Tyr-37B and Tyr-74 of monomer A, but also by the Phe-153 benzyl group and the 152-spur of the neighboring monomer B. Thus, binding of extended substrate chains is limited to about P5' (Fig. 7).

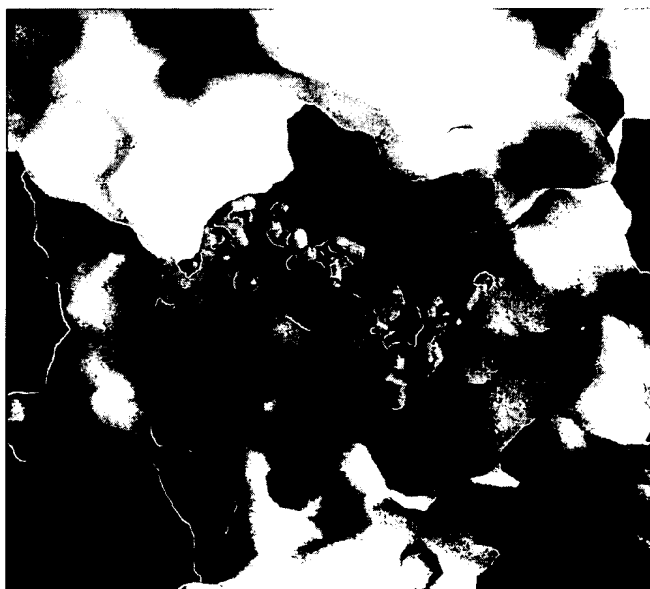


FIG. 7. View from the LDTI inhibitor (represented only by its reactive site loop P7 to P3') toward the active-site cleft. The P1 Lys residue is buried.

Toward the north, the 97-loop of monomer A borders the substrate binding region in a manner different from most other serine proteinases, and together with the side chains of Phe-94, Ala-97, and Gln-98 of monomer D forms a projecting "canopy." The S2 subsite underneath is open and larger than that of trypsin. The S3/S4 subsite above the Trp-215 indole moiety is fully blocked by the side chain of Gln-98 and the phenolic group of Tyr-95 provided by monomer D. Toward the west, however, the substrate-binding site is bordered exclusively by segments of the D-monomer, in particular the His-57 imidazole ring and segment 57-60. Thus, the active centers of monomers A and D (B and C) are spatially close (distance ≈ 23 Å for the A-D pair) to each other in the tryptase tetramer, rendering the tryptase tetramer suitable for the specific binding of bifunctional inhibitors with relatively short spacers.

The central pore of tryptase restricts the size of accessible substrates and inhibitors considerably. For larger proteins such as fibronectin and the zymogens of stromelysin-1 and urokinase-type plasminogen activator, the cleavage sites must be extended into the active sites. Docking experiments with C-terminally truncated prostromelysin-1 (45) and with single-chain tissue plasminogen activator (29) as a model for prourokinase show that the activation cleavage loops of these proproteinases must be extracted from their crystal structures to allow binding in the tryptase active center. More flexible peptides, in contrast, could easily thread through the pore of the tetramer to be processed or destroyed. Flexible polypeptide chains with two distant basic residues, as in "vasoactive intestinal peptide" (18), might even dock to adjacent active sites simultaneously to produce fragments of distinct length.

The active centers of the tryptase monomers are also largely inaccessible for macromolecular inhibitors. The only exception known is LDTI, an "atypical" Kazal-type inhibitor that is smaller than the classical members of this family (16). LDTI has been shown to bind to trypsin through its reactive-site loop (residues P4 to P4') in a canonical manner (17, 46). In the model of the complex with tryptase monomer A, the four N-terminal residues preceding this binding segment could bend toward the south (with respect to Figs. 3 and 7), leading to the juxtaposition of the basic Lys-11-Lys-12 amino terminus (with the suffix I identifying inhibitor residues) with the carboxylate groups of Asp-143 and Asp-147 of monomer A. Alternatively, the two Lys residues could interact with Asp-60B of molecule D. The involvement of such electrostatic interactions is supported by the deleterious effect of deletions and substitutions of these basic residues on the affinity of LDTI toward tryptase but not trypsin (17). The LDTI reactive-site loop, running from Cys-I14 (P5) to Pro-I22 (P4'; ovomucoid numbering), is relatively small compared with classical Kazal-type inhibitors, allowing good overall fit to the restricted substrate binding groove (Figs. 7 and 8a). Furthermore, its central helix is one turn shorter, so that it just fits into the central pore of the tetramer on canonical binding to the active site of monomer A with only a few narrow contacts of its molecular antipole, opposite to its reactive-site loop, with the 147-loop of monomer D. Docking of a second LDTI molecule is possible at the opposite active centers of either monomer B or monomer C (Fig. 8a). A slight collision between Cys-I56 and Gly-I28 of two bound LDTI molecules could be relieved by minor torsion in the proteinase-inhibitor interfaces, as observed for other canonically binding inhibitors such as eglin c (46). Any such torsion in the LDTI molecule bound to monomer A would impose an opposing torsion in the LDTI molecule bound to monomer B, facilitating such a relaxation. The simultaneous binding of two LDTI molecules to the tetramer is in good agreement with experimental results showing $\approx 50\%$ inhibition of the cleavage activity toward small chromogenic substrates by nanomolar LDTI concentrations (16). Modeling experiments with more elongated classical Kazal-type inhibitors or with the prototypical bovine pancre-

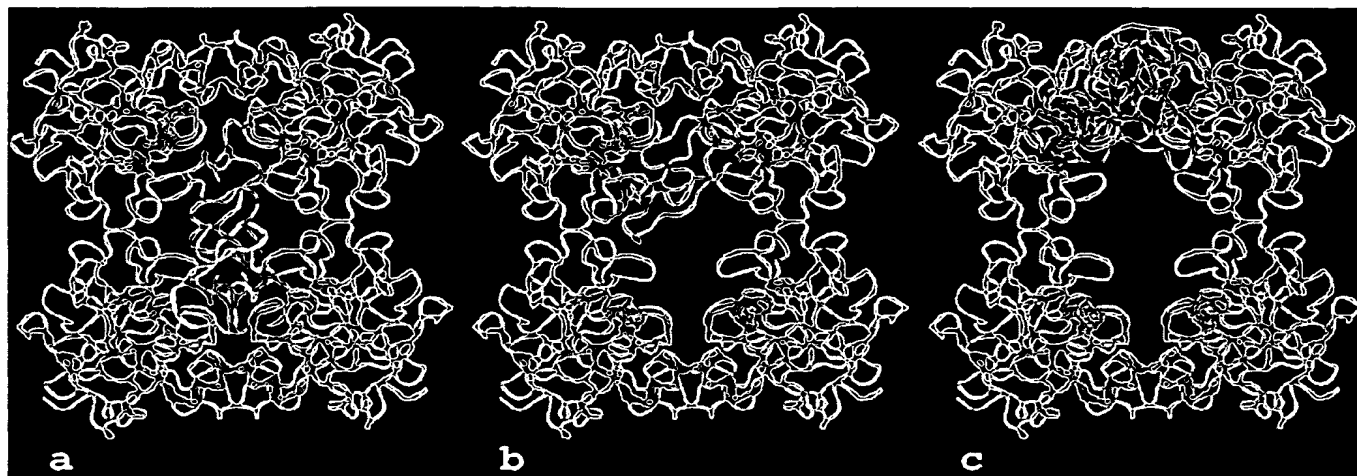


FIG. 8. Models of the interaction of the human tryptase tetramer with proteinaceous inhibitors. The tryptase tetramers are shown as green ribbons. An inhibitor molecule (blue) is modeled into the active site of monomer A by superposition of the proteinase moiety of known proteinase-inhibitor complexes to a tryptase monomer. For LDIT and BPTI the target proteinase was trypsin (17, 49), for MPI chymotrypsin (47). The active sites of the other tryptase monomers are occupied by APPA molecules (orange). Parts of the inhibitors clashing with the structure of tryptase (i.e., a distance smaller than 1.5 Å between the Ca-atoms of the respective molecules) are highlighted in red. (a) In addition to one molecule of the "atypical" Kazal-type inhibitor LDIT bound to the tryptase monomer A a second molecule (shown in pink and yellow) can bind to the active site of either monomer B or C. (b) Bovine pancreatic trypsin inhibitor (aprotinin). (c) Human mucous proteinase inhibitor bound to tryptase with its inhibitorily active second domain.

atic trypsin inhibitor indicate strong collisions of their distal pole segments with the neighboring monomers D and B, in particular with the 147-loops, explaining the observed inactivity of these inhibitors toward tryptase (Fig. 8b). The central portion of the two-domain mucous proteinase inhibitor (MPI = SLPI = HUSI-I) would clash with the A-D interface region of the tryptase tetramer if bound to the active site of monomer A (Fig. 8c) via its inhibitorily active second domain (47). Similarly, elafin (= SKALP), an inhibitor corresponding to the MPI second domain (48), should not be able to inhibit tryptase. The much larger plasma proteinase inhibitors are clearly far too bulky to fit into the narrow pore of the tryptase tetramer and gain access to one of the active centers.

CONCLUSION

In summary, the structure of the β II-tryptase tetramer has been identified based on the four crystallographically independent quasiidentical monomers and the analysis of their arrangement within the crystal packing. With its frame-like architecture and its active centers facing a narrow central pore, the resulting tryptase tetramer structure explains most of the distinct properties of the biologically active tryptase tetramer in solution. The unusual substrate specificity, with a preference for peptidergic substrates, and the resistance to proteinaceous inhibitors other than LDIT are both caused by the limited accessibility of the active sites within the narrow central pore. The tetramer can be stabilized by heparin glycosaminoglycan chains larger than ≈ 20 sugar residues, a length required to bridge the weaker of the two distinct monomer-monomer interfaces. The loss of enzymatic activity on dissociation of the tetramer is caused by stabilization by internal molecular groups of a zymogen-like rather than the active state. Finally, the knowledge of the structure of the active center of the monomer as well as of the distances between neighboring active sites allows the rational design of multifunctional inhibitors. Such inhibitors that bind to more than one active center will ideally have potentiated affinity, conferring selectivity for the tryptase tetramer. Such inhibitors will be valuable as pharmacological tools to probe the pathophysiological function(s) of tryptases *in vivo* and may have therapeutic potential against asthma and other mast-cell related disorders.

We are grateful to R. Huber and H. Fritz for their generous support. We thank D. Grosse and R. Mentele for their excellent help in crystallization and amino acid sequence analysis. This work was supported by Sonderforschungsbereich 469 of the University of Munich, the Deutsche Forschungsgemeinschaft (STU 161, BO 1279), the Fonds der Chemischen Industrie, and programs BIO4-CT98-0418 and TMR ERBFXCT 98-0193 of the European Union.

1. Miller, J. S., Westin, E. H. & Schwartz, L. B. (1989) *J. Clin. Invest.* **84**, 1188–1195.
2. Pallaoro, M., Fejzo, M. S., Shayesteh, L., Blount, J. L. & Caughey, G. H. (1999) *J. Biol. Chem.* **274**, 3355–3362.
3. Schwartz, L. B., Irani, A. M., Roller, K., Castells, M. C. & Schechter, N. M. (1987) *J. Immunol.* **138**, 2611–2615.
4. Xia, H. Z., Kepley, C. L., Sakai, K., Chelliah, J., Irani, A. M. & Schwartz, L. B. (1995) *J. Immunol.* **154**, 5472–5480.
5. Schwartz, L. B., Sakai, K., Bradford, T. R., Ren, S., Zweiman, B., Worobec, A. S. & Metcalfe, D. D. (1995) *J. Clin. Invest.* **96**, 2702–2710.
6. Sakai, K., Ren, S. & Schwartz, L. B. (1996) *J. Clin. Invest.* **97**, 988–995.
7. Caughey, G. H. (1997) *Am. J. Respir. Cell Mol. Biol.* **16**, 621–628.
8. Johnson, P. R. A., Ammit, A. J., Carlin, S. M., Armour, C. L., Caughey, G. H. & Black, J. L. (1997) *Eur. Respir. J.* **10**, 38–43.
9. Ricc, K. D., Tanaka, R. D., Katz, B. A., Numerof, R. P. & Moore, W. R. (1998) *Curr. Pharm. Des.* **4**, 381–396.
10. De Sanctis, G. T., Merchant, M., Beier, D. R., Dredge, R. D., Grobholz, J. K., Martin, T. R., Lander, E. S. & Drazen, J. M. (1995) *Nat. Genet.* **11**, 150–154.
11. Hunt, J. E., Stevens, R. L., Austen, K. F., Zhang, J., Xia, Z. & Ghildyal, N. (1996) *J. Biol. Chem.* **271**, 2851–2855.
12. Schwartz, L. B. (1994) *Methods Enzymol.* **244**, 88–100.
13. Caughey, G. H. (1995) *Mast Cell Proteases in Immunology and Biology* (Dekker, New York).
14. Ren, S., Sakai, K. & Schwartz, L. B. (1998) *J. Immunol.* **160**, 4561–4569.
15. Selwood, T., McCaslin, D. R. & Schechter, N. M. (1998) *Biochemistry* **37**, 13174–13183.
16. Sommerhoff, C. P., Söllner, C., Mentele, R., Piechottka, G. P., Auerswald, E. A. & Fritz, H. (1994) *Biol. Chem. Hoppe-Seyler* **375**, 685–694.
17. Stubbs, M. T., Morenweiser, R., Stürzebecher, J., Bauer, M., Bode, W., Huber, R., Piechottka, G. P., Matschiner, G., Sommerhoff, C. P., Fritz, H., et al. (1997) *J. Biol. Chem.* **272**, 19931–19937.

18. Tam, E. K. & Caughey, G. H. (1990) *Am. J. Respir. Cell Mol. Biol.* **3**, 27–32.
19. Gruber, B. L., Marchese, M. J., Suzuki, K., Schwartz, L. B., Okada, Y., Nagase, H. & Ramamurthy, N. S. (1989) *J. Clin. Invest.* **84**, 1657–1662.
20. Stack, M. S. & Johnson, D. A. (1994) *J. Biol. Chem.* **269**, 9416–9419.
21. Molino, M., Barnathan, E. S., Numerof, R., Clark, J., Dreyer, M., Cumashi, A., Hoxie, J. A., Schechter, N., Woolkalis, M. & Brass, L. F. (1997) *J. Biol. Chem.* **272**, 4043–4049.
22. Lohi, J., Harvima, I. & Keski-Oja, J. (1992) *J. Cell. Biochem.* **50**, 337–349.
23. Little, S. S. & Johnson, D. A. (1995) *Biochem. J.* **307**, 341–346.
24. Schwartz, L. B., Bradford, T. R., Littman, B. H. & Wintroub, B. U. (1985) *J. Immunol.* **135**, 2762–2767.
25. Pereira, P. J., Bergner, A., Macedo-Ribeiro, S., Huber, R., Matschiner, G., Fritz, H., Sommerhoff, C. P. & Bode, W. (1998) *Nature (London)* **392**, 306–311.
26. Blevins, R. A. & Tulinsky, A. (1985) *J. Biol. Chem.* **260**, 4264–4275.
27. Bode, W. & Schwager, P. (1975) *J. Mol. Biol.* **98**, 693–717.
28. Wang, D., Bode, W. & Huber, R. (1985) *J. Mol. Biol.* **185**, 595–624.
29. Renatus, M., Engh, R. A., Stubbs, M. T., Huber, R., Fischer, S., Kohnert, U. & Bode, W. (1997) *EMBO J.* **16**, 4797–4805.
30. Huber, R. & Bode, W. (1978) *Acc. Chem. Res.* **11**, 114–122.
31. Bode, W. (1979) *J. Mol. Biol.* **127**, 357–374.
32. Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. A. T. & Xuong, N. H. (1970) *Biochemistry* **9**, 1997–2009.
33. Bode, W., Fehlhämmer, H. & Huber, R. (1976) *J. Mol. Biol.* **106**, 325–335.
34. Hedstrom, L., Lin, T. Y. & Fast, W. (1996) *Biochemistry* **35**, 4515–4523.
35. Bode, W., Schwager, P. & Huber, R. (1978) *J. Mol. Biol.* **118**, 99–112.
36. Bolognesi, M., Gatti, G., Menagatti, E., Guarneri, M., Marquart, M., Papamokos, E. & Huber, R. (1982) *J. Mol. Biol.* **162**, 839–868.
37. Higashi, S. & Iwanaga, S. (1998) *Int. J. Hematol.* **67**, 229–241.
38. Pignol, D., Gaboriaud, C., Michon, T., Kerfelec, B., Chapus, C. & Fontecilla Camps, J. C. (1994) *EMBO J.* **13**, 1763–1771.
39. Banner, D. W., D'Arcy, A., Chene, C., Winkler, F. W., Guha, A., Konigsberg, W. H., Nemerson, Y. & Kirchhofer, D. (1996) *Nature (London)* **380**, 41–46.
40. Alter, S. C., Metcalfe, D. D., Bradford, T. R. & Schwartz, L. B. (1987) *Biochem. J.* **248**, 821–827.
41. Walter, J. & Bode, W. (1983) *Hoppe-Seyler's Z. Physiol. Chem.* **364**, 949–959.
42. Stürzebecher, J., Prasa, D. & Sommerhoff, C. P. (1992) *Biol. Chem. Hoppe-Seyler* **373**, 1025–1030.
43. Caughey, G. H., Raymond, W. W., Bacci, E., Lombardy, R. J. & Tidwell, R. R. (1993) *J. Pharmacol. Exp. Ther.* **264**, 676–682.
44. Katz, B. A., Clark, J. M., Finer Moore, J. S., Jenkins, T. E., Johnson, C. R., Ross, M. J., Luong, C., Moore, W. R. & Stroud, R. M. (1998) *Nature (London)* **391**, 608–612.
45. Becker, J. W., Marcy, A. I., Rokosz, L. L., Axel, M. G., Burbaum, J. J., Fitzgerald, P. M., Cameron, P. M., Esser, C. K., Hagmann, W. K., Hermes, J. D., et al. (1995) *Protein Sci.* **4**, 1966–1976.
46. Bode, W. & Huber, R. (1992) *Eur. J. Biochem.* **204**, 433–451.
47. Grütter, M. G., Fendrich, G., Huber, R. & Bode, W. (1988) *EMBO J.* **7**, 345–351.
48. Tsunemi, M., Matsuura, Y., Sakakibara, S. & Katsube, Y. (1996) *Biochemistry* **35**, 11570–11576.
49. Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J. & Steigemann, W. (1974) *J. Mol. Biol.* **89**, 73–101.

Exhibit 35



The Three-Dimensional Structure of Asn¹⁰² Mutant of Trypsin: Role of Asp¹⁰² in Serine Protease Catalysis

S. Sprang; T. Standing; R. J. Fletterick; R. M. Stroud; J. Finer-Moore; N-H. Xuong; R. Hamlin; W. J. Rutter; C. S. Craik

Science, New Series, Vol. 237, No. 4817 (Aug. 21, 1987), 905-909.

Stable URL:

<http://links.jstor.org/sici?sici=0036-8075%2819870821%293%3A237%3A4817%3C905%3ATTSOAM%3E2.0.CO%3B2-E>

Science is currently published by American Association for the Advancement of Science.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/aaas.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

<http://www.jstor.org/>
Thu Sep 9 18:09:41 2004

15. Hosts were mounted in 1.34- and 1.00-mm diameter holes in white plastic squares (2 by 2 cm). Each wasp was allowed to complete examination and oviposition. The wasps were observed individually to prevent repeated parasitization of the same host. Trials in which the wasp left the host before completing oviposition were rejected.
16. Mean \pm SD was used throughout. Statistical significance was determined by *t* tests.
17. S. E. Flanders, *Proc. Pac. Entomol.* 11, 175 (1935).
18. Head length was measured from the medial ocellus to the tip of the closed mandibles by using an ocular

- micrometer. Wasps differed significantly in mean head length between large and small treatment groups ($P < 0.001$).
19. Single hosts were mounted on white cardboard squares (2 by 2 cm) with gum arabic. After host examination was completed, wasps were observed as in (15).
20. Measurements made from films of the initial transit demonstrate a significant linear relation between wasp body length and stride length [slope, 0.58 ± 0.064 (SE); $n = 15$, $P < 0.01$].
21. Wasps were observed on single hosts mounted on

cardboard cards with gum arabic. Only wasps that completed their host examination and began ovipositing were included in the data. For details of methods and results, see J. M. Schmidt and J. J. B. Smith [*J. Exp. Biol.* 129, 151 (1987)].

22. Lepidoptera: Gelechiidae.
23. We thank R. Tanner and R. Chaplinsky for technical assistance. The Natural Sciences and Engineering Research Council of Canada provided financial support.

4 March 1987; accepted 1 June 1987

The Three-Dimensional Structure of Asn¹⁰² Mutant of Trypsin: Role of Asp¹⁰² in Serine Protease Catalysis

S. SPRANG,* T. STANDING, R. J. FLETTERICK, R. M. STROUD, J. FINER-MOORE, N.-H. XUONG, R. HAMLIN, W. J. RUTTER, C. S. CRAIK

The structure of the Asn¹⁰² mutant of trypsin was determined in order to distinguish whether the reduced activity of the mutant at neutral pH results from an altered active site conformation or from an inability to stabilize a positive charge on the active site histidine. The active site structure of the Asn¹⁰² mutant of trypsin is identical to the native enzyme with respect to the specificity pocket, the oxyanion hole, and the orientation of the nucleophilic serine. The observed decrease in rate results from the loss of nucleophilicity of the active site serine. This decreased nucleophilicity may result from stabilization of a His⁵⁷ tautomer that is unable to accept the serine hydroxyl proton.

THROUGHOUT THE DIVERSE FAMILY of serine proteases, the three residues implicated in the bond breaking and making events of protease catalysis, His⁵⁷, Asp¹⁰², and Ser¹⁹⁵ (chymotrypsin numbering system) are conserved. The spatial relation among these residues is virtually equivalent in the three-dimensional structures of all serine proteases studied. The catalytic roles of Ser¹⁹⁵ and His⁵⁷ are firmly established (1). The substrate (ester or amide) carbonyl carbon undergoes a nucleophilic attack by the hydroxyl group of Ser¹⁹⁵, which leads to the formation of an acyl enzyme intermediate. His⁵⁷ functions as a catalytic base by assisting in the transfer of a proton from the serine hydroxyl to the substrate leaving group. The role of Asp¹⁰² has not yet been defined. The three functions proposed for this residue are: (i) stabilizing the His⁵⁷ conformation that is required for catalysis (2), (ii) stabilizing the

appropriate His⁵⁷ tautomer (2), and (iii) stabilizing the positively charged histidine that forms during the reaction (3). The proposed functions were tested with a ge-

netically engineered mutant of the anionic isozyme of rat trypsin that was constructed by replacing Asp¹⁰² with an asparagine (4), designated here as D 102 N trypsin, where D is Asp and N is Asn.

The activity of D 102 N trypsin has been studied as a function of pH (4). The activity of this mutant enzyme toward a variety of substrates is reduced by four orders of magnitude relative to trypsin between pH 7 and pH 9, where the latter is optimally active. The Michaelis constant, K_m , of the mutant enzyme is virtually unaffected (4). This raises the question of whether the chemical properties of the asparagine itself or the conformational differences in the enzyme are responsible for the loss of activity in D 102 N trypsin. To address this point, we describe the three-dimensional structure of D 102 N trypsin at both pH 6 and pH 8.

Orthorhombic crystals (space group $P2_12_12_1$) of rat D 102 N trypsin grown at pH 6 in the presence of benzamidine were

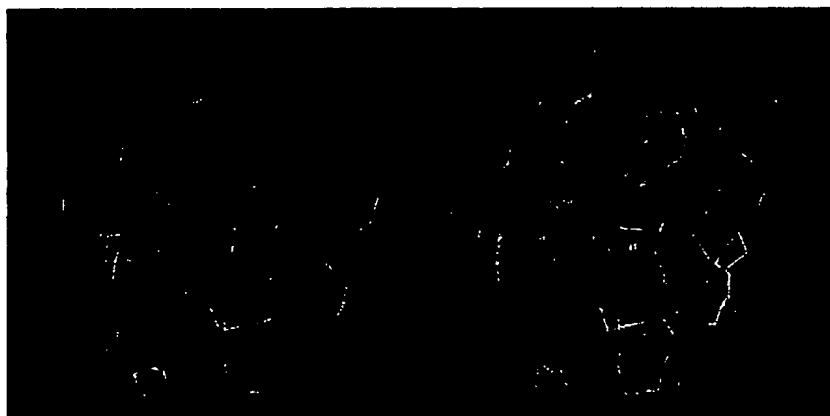


Fig. 1. An α -carbon diagram (stereoscopic) of anionic rat D 102 N trypsin at pH 6 (9-12) (green) is superimposed on bovine trypsin (blue). Residues in rat trypsin (12) that differ in side-chain type from corresponding residues in the bovine sequence (25) are highlighted in red here. Side-chain positions for residues Asn¹⁰², His⁵⁷, and Ser¹⁹⁵ are also shown in red. The root-mean-square (rms) difference in position between corresponding atoms of D 102 N rat trypsin in the crystals grown at pH 6 and bovine trypsin (13, 26) after least-squares superposition is 0.47 Å for all main-chain atoms and 0.67 Å for all side-chain atoms. Values quoted are the average of those obtained for molecules 1 and 2 in the asymmetric unit of the D 102 N trypsin crystals grown at pH 6. The computed rms distance may be an underestimate of the true differences in the two structures because of the use of bovine trypsin as the initial phasing model. The rms difference after superposition between all atoms of the two molecules in the asymmetric unit is 0.21 Å. The rms deviation between the main-chain atoms of the pH 6 and pH 8 crystal forms of D 102 N trypsin is 0.25 Å.

S. Sprang, T. Standing, R. J. Fletterick, R. M. Stroud, J. Finer-Moore, Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143.

N.-H. Xuong and R. Hamlin, Department of Physics, University of California, San Diego, La Jolla, CA 92093. W. J. Rutter, Hormone Research Institute, University of California, San Francisco, San Francisco, CA 94143.

C. S. Craik, Department of Biochemistry and Biophysics and Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94143.

*Present address: Howard Hughes Medical Institute, University of Texas, Dallas, TX 75235.

obtained by vapor diffusion against polyethylene glycol (Figs. 1 and 2, top). Diffraction data were measured to 2.3 Å resolution with monochromatic copper K α radiation and the crystal cooled to 4°C on a multiwire area detector with the procedures described by Xuong *et al.* (5) (Table 1). A cubic crystal form (space group *I*23) was obtained at pH 8 by vapor diffusion against magnesium sulfate. Diffraction data for this form were recorded to 2.8 Å resolution with monochromatic copper K α radiation on a diffractometer (7) (Table 1 and Fig. 2, middle). Both crystal structures were determined by molecular replacement methods (8) and refined by stereochemically restrained minimization of the differences between observed and computed structure amplitudes (6, 9–12) (Table 1 and Figs. 1 and 2).

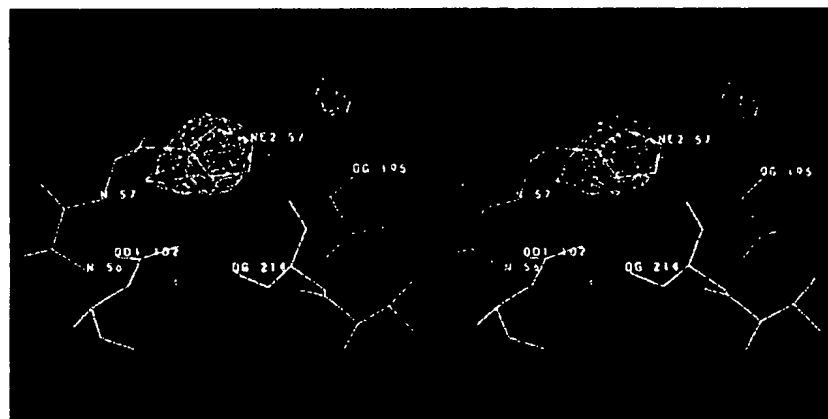
The tertiary structures of the mutant rat anionic trypsin at both pH 6 and 8 are essentially identical to that of the bovine enzyme (7, 13). The largest differences between the enzymes from rat and cow are localized to four segments in the NH₂ terminal domain, all outside the β core, where deviations between corresponding main chain atoms exceed 1.0 Å (Fig. 1). The structural similarity between D 102 N tryp-

sin and bovine trypsin is quite high in the neighborhood of the active site; no significant differences in the relative positions (<0.3 Å) (Table 2) or relative thermal factors are observed for Asn¹⁰², Ser¹⁹⁵, or the oxyanion binding site (14); that is, the main-chain amide groups of residues 193 and 195. The only exception occurs in crystals grown at pH 6, where the side chain of His⁵⁷ is statistically disordered (Fig. 2, top) (11, 12), and is partitioned between the gauche conformation observed in native trypsin and an alternative trans conformation, in which the imidazole side chain is

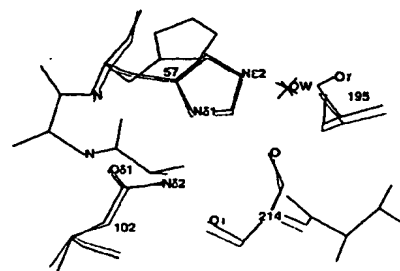
displaced from the active site toward the solvent. Only the native gauche His⁵⁷ conformation is observed in crystals grown at pH 8. Unless otherwise stated, all references to His⁵⁷ in the following discussion refer to the native conformer.

In both the pH 8 and pH 6 crystal forms, Asn¹⁰² is superimposable within experimental error with Asp¹⁰² of the bovine enzyme (Fig. 2). In trypsin, one of the carboxylate oxygen atoms of Asp¹⁰² accepts hydrogen bonds from the main-chain amide groups of residues 56 and 57, and the second accepts hydrogen bonds from both the N δ 1 atom of

Fig. 2. (Top) The difference Fourier map ($F_{\text{obs}} - F_{\text{calc}}$) at the catalytic site of D 102 N rat trypsin at pH 6. The side-chain atoms of His⁵⁷ were omitted from the calculated structure factors and phases. The trans and gauche conformations of the histidine side chain related by χ^1 torsional differences of 70° are superimposed on the electron density. The difference electron density is shown at a contour level of 0.2 electron per cubic angstrom. The map extends over all atoms shown in the figure. No negative density is present in this region at the 0.2 electron per cubic angstrom level. Two lobes of flat, ellipsoidal density are evident, both continuous with the density corresponding to the C β atom of His⁵⁷. The peaks are of unequal magnitude; the stronger peak is located within the active site between the side chains of Asn¹⁰² and Ser¹⁹⁵ at a position coincident with His⁵⁷ in the structures of bovine trypsin, and the second weaker peak is outside of the active site pocket. The shape of both lobes of density and their proximity to the C β atom of His⁵⁷ rules out the assignment of either peak to ordered solvent. **(Middle)** A difference Fourier map ($F_{\text{obs}} - F_{\text{calc}}$) showing the catalytic site of D 102 N trypsin from crystals grown at pH 8. The side-chain atoms of His⁵⁷ were omitted from the calculated structure factors and phases. At this pH, only the gauche conformer for His⁵⁷ is observed in the difference electron density. The histidine conformation is almost identical to that observed in bovine trypsin–benzamidine complex (7). The structure of D 102 N trypsin at pH 8 was determined by molecular replacement, using the refined structure at pH 6 as a search model. The side-chain atoms of Asn¹⁰², His⁵⁷, and Ser¹⁹⁵ as well as solvent, benzamidine, and calcium ion atoms were omitted from this model. The rotation function produced only one significant peak and was evaluated with all data to 2.8 Å and an



integration radius from 4.0 to 16 Å. The *R* factor at the correct translation position was 0.35. A difference Fourier map computed with phases from the molecular replacement solution revealed the positions of the omitted side chains, calcium ion, and benzamidine molecule. These were included in the phasing model and the structure was subjected to 23 cycles of stereochemically restrained crystallographic refinement (Table 1) (6). **(Bottom)** The bovine trypsin structure (thin lines) is superimposed on that of D 102 N rat trypsin crystallized at pH 6.0 (thick lines). Both conformers of His⁵⁷ in D 102 N rat trypsin are shown.



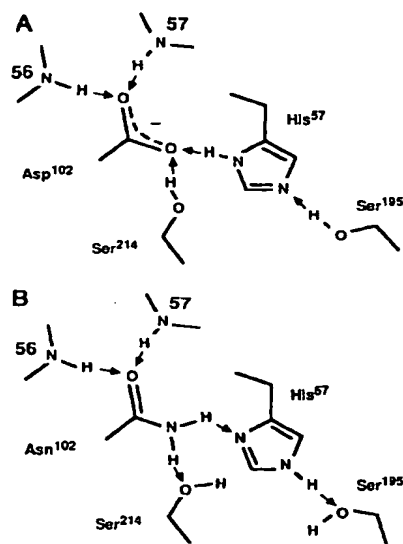


Fig. 3. (A) In the hydrogen bond network found in D 102 N trypsin above neutral pH, His⁵⁷ is unable to accept a proton from Ser¹⁹⁵ Oδ. The orientation of the hydrogen bond between His⁵⁷ and Ser¹⁹⁵ is the reverse of that observed in the bovine trypsin-benzamidine structure (7). (B) In the hydrogen bond network of wild-type trypsin, His⁵⁷ is an acceptor for the proton from Ser¹⁹⁵.

Table 1. Crystal and diffraction data for D 102 N trypsin. The diffraction data for the crystals grown at pH 6 were collected with an area detector, whereas the data for the crystals grown at pH 8 were collected with a diffractometer.

Diffraction data	Crystal form	
	pH 6	pH 8
Crystal data		
Space group	P2 ₁ 2 ₁ 2 ₁	I23
Cell dimensions (Å)	a = 40.4 b = 92.0 c = 127.4	a = 124.4
Molecules per asymmetric unit	2	1
Diffraction data		
Resolution (Å)	2.3	2.8
Total observations	90,000	5,000
Unique observations	22,000	4,500
R _{sym} ^a	0.05	
Refinement results		
R _{cryst} [†]	0.16	0.21
Resolution (Å)	6.0–2.3	8.0–2.8
rms difference (bond) (Å) [‡]	0.03	0.03
rms difference (angle) (Å) [‡]	0.05	0.05

^aAgreement between symmetry-related structure-factor magnitudes $R = (\sum_i \sum_j |(F_{hi}) - (F_{hj})|) / (\sum_i \sum_j F_{hi})$

where (F_{hi}) is the mean structure factor magnitude of the i observations of reflections that are related to the Bragg index h . [†]Agreement between the observed (F_{obs}) and calculated (F_{calc}) structure factor magnitudes $R_{cryst} = (\sum_i |F_{obs,i} - F_{calc,i}|) / (\sum_i F_{obs,i})$

[‡]Root-mean-square deviation between the ideal and refined bond distances and angle distances.

His⁵⁷ and the Oγ atom of Ser²¹⁴ (Table 2 and Fig. 3). In D 102 N trypsin, there are two chemically distinct conformations possible for Asn¹⁰². In one of these the Nδ2 group of Asn¹⁰² would be oriented toward the main-chain amide groups of residues 56 and 57. Since the asparagine amide group cannot form a hydrogen bond with the main-chain amides in this orientation, they could approach no closer than the sum of their van der Waals radii (>3.4 Å).

The alternative conformation is related to the first by a rotation of 180° about the Cβ–Cγ bond. In this case, the Oδ1 atom of asparagine could accept hydrogen bonds from the main-chain amide groups, whereas the Nδ2 atom could accept hydrogen bonds from the His⁵⁷ imidazole and Ser²¹⁴ hydroxyl groups. The two conformations can be distinguished by the observed distances between the main-chain amides of residues 56 and 57 and the nearest atom of the Asn¹⁰² side chain. The interatomic distances in the present model (15, 16) support the assignment of the tautomeric form shown in Fig. 3A. One of the Asn¹⁰² amide atoms is located 2.6 Å from the amide nitrogen of residue 56 and 3.1 Å from the amide of residue 57. This atom of the Asn¹⁰² side chain could then be involved in hydrogen bonds with these two amides and would thus be identified as Oδ1. Asn¹⁰² Nδ2 would therefore be a hydrogen bond donor to both the Nδ1 of His⁵⁷ and the Oδ of Ser²¹⁴. Asp¹⁰² accepts hydrogen bonds from both of these residues in bovine trypsin.

In the proposed crystallographic model, Asn¹⁰² can only serve as a hydrogen bond donor to His⁵⁷; the polarity of the hydrogen bond network involving His⁵⁷, residue 102, and Ser¹⁹⁵ is reversed in the mutant enzyme with respect to that in bovine trypsin (Fig. 3). For values of pH greater than the pK_a of the imidazole (K_a is the ionization constant), the monoprotonated tautomer must be protonated at Ne2 since it serves as a hydrogen bond acceptor from Asn¹⁰² at Nδ1. In contrast to trypsin, the Ne2 of

His⁵⁷ in the mutant enzyme is a potential hydrogen bond donor to the Oγ of Ser¹⁹⁵. Thus His⁵⁷ cannot act as a general base in transferring a proton from Ser¹⁹⁵ and this probably accounts for the diminished activity of D 102 N trypsin near neutral pH. For trypsin above neutral pH, where the enzyme becomes active, His⁵⁷ is protonated at Nδ1 (17). Therefore, the presence of a negatively charged Asp¹⁰² maintains the unprotonated Ne2 with a lone pair of electrons as the general base catalyst for transfer of the proton from the Oγ of Ser¹⁹⁵ to the leaving group.

A difference Fourier map (Fig. 2, top) for the crystals grown at pH 6 was computed with the histidine omitted from the calculated phases and structure factors, revealing two sites for the side chain (11, 12). In one of these, the Cβ–Cγ bond is trans to Cα–N, and the imidazole is rotated from the catalytic site. The trans His⁵⁷ conformer does not form a hydrogen bond with Asn¹⁰² or Ser¹⁹⁵ but rather is in contact with a solvent water molecule at the surface of the enzyme (Table 2). The alternative position is nearly gauche and similar to the His⁵⁷ conformation in bovine trypsin and D 102 N trypsin crystallized at pH 8 (Fig. 2, bottom).

Integration of the difference electron density indicates that the occupancy ratio of the gauche to trans isomers is approximately 2 to 1 (Fig. 2, top) (11, 12). A difference map computed with phases derived from all of the atoms in the refined model reveals residual positive electron density in the vicinity of the Cε1 of His⁵⁷ (gauche), and may correspond to a partially occupied solvent water which is present in the active site pocket when His⁵⁷ is displaced (trans).

The displacement of His⁵⁷ from the active site of D 102 N trypsin below neutral pH is probably a consequence of steric conflicts between the protonated Nδ1 atom of His and the proton on the Nδ2 of Asn. D 102 N trypsin, like its natural homolog, is crystallized only in the presence of the substrate analog benzamidine, and there are no appar-

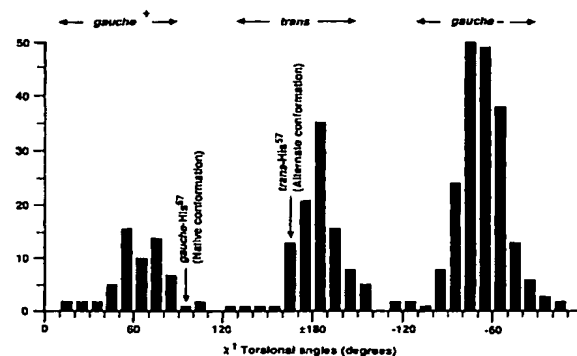


Fig. 4. A histogram showing the χ^1 torsion angles of 353 histidines found in 53 protein structures refined to greater than 2.0 Å resolution (11, 26). The χ^1 angle of 92° gauche observed in His⁵⁷ of bovine trypsin is rare. Angle values are trimodally distributed about +60°, 180°, and -60°. The trans conformer that occurs at pH 6 in D 102 N rat trypsin is more frequently observed.

Table 2. Conformational and stereochemical data for active site residues in bovine and D 102 N trypsin. Values for the two molecules in the asymmetric unit of D 102 N trypsin grown at pH 6 are averaged. Distances are not given for the 2.8 Å resolution crystals grown at pH 8. The wild-type coordinates are from the bovine trypsin-benzamidine crystal structure (7).

Residue	Atoms	Conformational angles (degrees)		Hydrogen bond distance (Å)	
		Asn ¹⁰²	Wild type	Asn ¹⁰²	Wild type
His ⁵⁷ (gauche)	N-Cα-Cβ-Cγ	84	92		
His ⁵⁷ (trans)		157			
His ⁵⁷ (gauche)	Cα-Cβ-Cγ-N81	-96	-100		
His ⁵⁷ (trans)		-93			
Ser ¹⁹⁵	N-Cα-Cβ-Oγ	-59	-77		
His ⁵⁷ (gauche)	N81-Asn/Asp ¹⁰² N/O82			2.8	2.7
His ⁵⁷ (gauche)	Ne2-Ser ¹⁹⁵ Oγ2			3.2	3.0
His ⁵⁷ (gauche)	Ne2-H ₂ O ²⁹³ O			3.0	
Asn ¹⁰² /Asp ¹⁰²	O81-Ala ⁵⁶ N			2.6	2.9
Asn ¹⁰² /Asp ¹⁰²	O81-His ⁵⁷ N			3.1	2.8
Asn ¹⁰² /Asp ¹⁰²	N/O82-Ser ¹¹⁴ Oγ			2.7	2.8
Ser ¹⁹⁵	Oγ-H ₂ O ⁷¹⁰ O			2.9	3.0

ent steric conflicts between His⁵⁷ and other residues in the catalytic site. However, even in trypsin, the native gauche conformation of His⁵⁷ imidazole may be energetically unfavored and require hydrogen bond stabilization by Asp¹⁰². A survey of the χ^1 angles of His side chains in refined protein structures (Fig. 4) shows that the conformation found in bovine trypsin is uncommon. Steric hindrance arises as a result of close contacts between the Cγ and C82 imidazole atoms and the main-chain carbonyl carbon [contact distances of 3.0 Å and 3.2 Å, respectively, are measured from the coordinates of bovine trypsin (7)]. Nevertheless, His⁵⁷ is well ordered in crystals of native trypsin (13, 17) and tritium exchange measurements indicate that expulsion of His⁵⁷ from the active site pocket occurs in solution with a frequency of less than 1 in 50 over the pH range 1.5 to 9 (18). Displacement of His⁵⁷ from the gauche conformation in serine protease crystals has so far been seen to occur as a result of steric conflict in covalent intermediates formed with certain substrate analogs (19, 20) or as a result of the introduction of heavy metals into the active site (21, 22). In native trypsin, the histidine conformation is stabilized by a hydrogen bond between the N81 atom of His and the carboxylate oxygen atom of Asp¹⁰².

In D 102 N trypsin, the conformation of His⁵⁷ appears to be linked to its protonation state. In the monoprotonated imidazole tautomer that predominates above neutral pH, the N81 atom of His can accept a hydrogen bond from N82 of Asn¹⁰². Protonation at the histidine N81 at the lower pH results in the loss of this hydrogen bond and possibly also steric conflict with the N82 of Asn¹⁰². The imidazole is then free to rotate to the more favored trans conformation, away from the catalytic site. Orthorhombic crys-

tals of D 102 N trypsin are grown near the pK_a of histidine, and thus the statistically disordered histidine side chain may reflect an equilibrium distribution of mono (gauche) and diprotonated (trans) forms of the His⁵⁷ imidazole. The variant D 102 N trypsin is able to react with the active site titrant tosyl-L-lysine chloromethyl ketone (TLCK) at 20 to 70% of the rate observed for trypsin from pH 7.2 to 8.7 (4), which suggests that as in the pH 8 crystals, a substantial proportion of D 102 N trypsin molecules in solution contain His⁵⁷ in the native gauche conformation.

As a result of the substitution of Asn for Asp¹⁰², the mutant trypsin reacts with diisopropylfluorophosphate (DFP), a reagent that specifically titrates the Ser¹⁹⁵ nucleophile, 10⁴ times more slowly than with trypsin (4). The decreased Ser¹⁹⁵ nucleophilicity in D 102 N trypsin probably results from the lack of a base in the active site to accept the serine hydroxyl proton. His⁵⁷ does not act as a base in this mutant because it exists in the incorrect tautomer. While the tautomeric form of His⁵⁷ is changed in D 102 N trypsin, the oxyanion binding site (14)—the main-chain amide groups of residues 193 and 195—is unaltered. The reduced activity of the mutant thus gives an upper limit to the contribution of transition state binding alone to the reaction rate. Trypsin normally accelerates the rate of DFP hydrolysis by a factor of 10⁸ (20). Our results suggest that a factor of 10⁴ in rate enhancement may derive from the stabilization and orientation of the lone pair on the Ne2 atom of His⁵⁷. The remaining factor of 10⁴ can presumably be ascribed to orientation of the nucleophile (Ser¹⁹⁵) and transition state binding. Under alkaline conditions (pH > 10), the rate of catalysis by the mutant approaches 10% of that of the native

enzyme (4) through an altered mechanism in which base catalysis appears to be provided by solvent hydroxide. In trypsinogen, the situation is reversed; His⁵⁷ is correctly oriented, but the oxyanion binding site is not properly formed to stabilize the transition state (21), even after irreversible binding of the transition state analog DFP (23). The reaction rate toward DFP is also reduced by a factor of ~10⁴ relative to trypsin (20), which again ascribes an upper limit of 10⁴ rate acceleration to transition state binding. Catalytic rate enhancement by serine proteases is thus partitioned almost equally between (i) orientation and stabilization of the enzyme base His⁵⁷ and (ii) the correctly oriented serine nucleophile and transition state binding site. Studies of D 102 N trypsin indicate that the Asp¹⁰² residue plays a critical role in the first of these processes, perhaps electronically with His⁵⁷ (24), and structurally, by providing hydrogen bond stabilization of the functional tautomer and thus maintaining its correct orientation within the catalytic site.

REFERENCES AND NOTES

- G. Dixon, S. Go, H. Neurath, *Biochim. Biophys. Acta* 19, 193 (1956); E. Shaw, M. Mares-Guia, W. Cohen, *Biochemistry* 4, 2219 (1965).
- W. W. Bachovchin, *Biochemistry* 25, 7751 (1986).
- A. R. Fersht and J. Sperling, *J. Mol. Biol.* 74, 137 (1973); A. R. Fersht, *Enzyme Structure and Mechanism* (Freeman, New York, ed. 1, 1977).
- C. S. Craik et al., *Science* 237, 909 (1987).
- N. H. Xuong, S. T. Freer, R. Hamlin, C. Nielsen, W. Vernon, *Acta Crystallogr.* A34, 289 (1978); R. Hamlin et al., *J. Appl. Crystallogr.* 14, 85 (1981); A. J. Howard, C. Nielsen, N. H. Xuong, *Methods Enzymol.* 114, 452 (1985).
- W. A. Hendrickson and J. H. Konert, in *Computing in Crystallography*, R. Diamond, S. Rameshnan, K. Venkatesan, Eds. (Indian Academy of Sciences, Bangalore, 1980), p. 13.01. The program PROLSQ written by Hendrickson and Konert was used for refinement. The modification added to PROLSQ by B. Finzel was used to decrease computation time.
- J. L. Chambers and R. M. Stroud, *Acta Crystallogr.* B33, 1824 (1977); *ibid.* B35, 1861 (1979); M. Krieger, L. M. Kay, R. M. Stroud, *J. Mol. Biol.* 83, 209 (1974).
- M. G. Rossmann, Ed., *The Molecular Replacement Method* (Gordon & Breach, New York, 1972); R. A. Crowther, *ibid.*, p. 173.
- The structure of D 102 N rat trypsin at pH 6 was determined by molecular replacement methods (8) by using the atomic coordinates of bovine trypsin-benzamidine (7) as a search set. The coordinates were modified by removal of all side-chain atoms for positions at which rat and bovine trypsin differ in amino acid sequence, as well as those of His⁵⁷ and Ser¹⁹⁵. Coordinates for solvent (benzamidine) and the bound calcium ion were excluded. The crystals grown at pH 6 exhibit pseudotranslational symmetry such that the unit cell comprises a b axis repeat of two P2₁2₁2₁ subcells related by a translation of $b/2$. As a consequence, reflections with b odd for $a \approx d \approx 3.0$ Å are systematically weak or absent. The relative rotation of the search coordinates with respect to the rat trypsin unit cell was determined by using the fast rotation function developed by Crowther (8). The correct solution was found with data to 3.0 Å resolution and an integration radius from 4.5 Å to 16.0 Å. The position of the rotated search model in the D 102 N trypsin unit cell was found by an R-factor search (with a program obtained from E. Dodson and P. Evans), which gave an R factor of 0.43. The position was refined by least

squares with the computer program CORELS (10). The positional parameters of individual atoms were then refined subject to stereochemical restraints by using the subcell data (6). The positions of missing side-chain atoms and those of the benzamidine and calcium were determined from the subcell difference electron density map computed from the refined model. A model of the full crystallographic asymmetric unit in the correct $P2_12_12_1$ unit cell was then constructed by adding a replicate of the trypsin molecule translated by 46 Å along the b and 32 Å along c . The full model was refined in three stages. In each stage the model was refit to a difference Fourier map computed with the coefficients ($2F_{\text{obs}} - F_{\text{calc}}$). Strong peaks in the electron density in positions consistent with hydrogen bond contacts to the protein or other established solvent positions were included in the model as ordered solvent. Next, the positional and thermal parameters of all atoms were refined by iterations of restrained crystallographic least squares, with data in the resolution range $6 \text{ Å} \leq d \leq 2.3 \text{ Å}$. Refinement was stopped when further cycles failed to reduce the crystallographic R factor and when the mean shift in coordinate positions was less than 0.05 Å. Refined coordinates were then used to compute phases for a new electron map to be used in the next stage of manual refitting. After the third stage (R factor = 0.18), examination of the electron density failed to reveal errors or ambiguity in main- or side-chain positions, although the side chains of six residues located at the surface of the molecules were disordered and could not be defined. Up to this point, side-chain atoms for His³⁷, Asn¹⁰², or Ser¹⁹⁵ had been excluded from the model. A difference electron density map ($F_{\text{obs}} - F_{\text{calc}}$) revealed strong and well-ordered density for the Asn¹⁰² and Ser¹⁹⁵, but the His³⁷ residue appeared to be statistically disordered (Fig. 2, top) (11).

10. J. L. Sussman, S. R. Holbrook, G. M. Church, S. H. Kim, *Acta Crystallogr.* A32, 311 (1976).

11. The possibility that one or other of the peaks are artifactual was tested by independent refinement of two alternative models: one with His³⁷ fit to the stronger, internal density and the second with His³⁷ fit to the external density. In each model the His³⁷ atoms were assigned full occupancy and side-chain positions for Asn¹⁰² and Ser¹⁹⁵ were included. Each model was subjected to restrained crystallographic refinement by varying the thermal and positional parameters of all atoms. Subsequently, a difference Fourier map ($F_{\text{obs}} - F_{\text{calc}}$) was computed for each model with the use of the refined positional and thermal parameters for all of the atoms in the respective models. In both cases, residual electron density appeared at the alternative histidine site. Again, the observed density peaks were contiguous with the C β atom of His³⁷ and thus could not be interpreted as ordered water molecules. The relative occupancy of the two histidine positions and the total occupancy of both positions relative to other histidine side chains was estimated by integration of difference electron density at all of the histidine side-chain positions in one of the trypsin molecules in the asymmetric unit. The difference Fourier map ($F_{\text{obs}} - F_{\text{calc}}$) used in the integration was computed from a model in which the side-chain atoms of all four histidine residues (at sequence positions 40, 57, 70, and 87) were removed from the coordinate set of one molecule. Integration was performed manually by summing over all grid points within 2.0 Å of histidine atomic positions that had electron density at least one standard deviation greater than the background density. After normalization the apparent relative integrated difference densities at the histidine side-chain positions were: His⁴⁰, 0.87; His⁵⁷, 0.60; His⁷⁰, 0.79; and His⁸⁷, 1.0. All but His³⁷ are well ordered, so the range in integrated densities reflects thermal motion and experimental error. The sum of the density over the two His³⁷ side-chain sites is lower than the mean density of the well-ordered histidine side chains, but is consistent with the high B factors of His³⁷ atoms at both positions. The relative occupancy of the alternative His³⁷ positions was estimated by integrating the difference density at the N δ 1 and C ϵ 1 atoms of the gauche conformer and the C δ 2 and N ϵ 2 atoms of the trans conformer and by taking the ratio of the

integrated densities for the two positions. The remaining histidine atoms were not included in the integration because the resolution of the data set did not allow the densities of the two conformers to be resolved at those positions.

Final refined positional and thermal parameters for both trans and gauche conformers were determined by refining an atomic model in which both conformers were simultaneously included. Side-chain atoms of the gauche conformer were assigned occupancies of 0.67 and atoms of the trans isomer were assigned occupancies of 0.33 based on the estimate derived from the integration described above (12). After three final cycles of refinement of all thermal and positional parameters of both trypsin monomers in the asymmetric unit, the crystallographic R factor was 0.161.

12. A modified version of PROTEIN (obtained from J. Smith) does not generate restraints between alternate side-chain positions of a statistically disordered residue. This allows refinement of two conformations of an amino acid simultaneously.

13. W. Bode and P. Schwager, *J. Mol. Biol.* 98, 693 (1975).

14. R. Henderson, *ibid.* 54, 341 (1970).

15. An upper estimate of the mean error in atomic position is 0.25 Å. It was obtained by an analysis of the variation of crystallographic R factor as a function of resolution (16).

16. V. Luzzati, *Acta Crystallogr.* 6, 142 (1953).

17. A. A. Kossiakoff and S. A. Spencer, *Biochemistry* 20,

6462 (1981).

18. M. Krieger *et al.*, *ibid.* 15, 3458 (1976).

19. M. N. G. James, A. R. Sielecki, G. D. Brayer, L. T. Delbacc, C. A. Bauer, *J. Mol. Biol.* 144, 43 (1980).

20. P. H. Morgan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 69, 3312 (1972).

21. A. A. Kossiakoff *et al.*, *Biochemistry* 16, 654 (1977); H. Fehllhammer, W. Bode, R. Huber, *J. Mol. Biol.* 111, 415 (1977).

22. J. L. Chambers *et al.*, *Biochem. Biophys. Res. Commun.* 59, 70 (1974).

23. M. O. Jones and R. M. Stroud, *Biochemistry*, in press.

24. D. M. Blow *et al.*, *Nature (London)* 221, 337 (1969).

25. C. S. Craik *et al.*, *J. Biol. Chem.* 259, 14255 (1984).

26. The coordinates were obtained from the Protein Data Bank at Brookhaven National Laboratory.

27. We thank J. Sadowsky, C. Neilsen, and E. Goldsmith for assistance with Area Detector data collection and processing and B. Montfort for assistance with crystallographic refinement calculations. We gratefully acknowledge grant support from NIH: AM31507 to S.R.S., GM24485 to R.M.S., and AM26081 to R.J.F.; from NSF: DMB8608086 to C.S.C. and PCM830610 to W.J.R.; a Bristol Meyer grant of Research Corporation and a CCRC grant to C.S.C. The coordinates of the D 102 N trypsin structure at pH 6 have been submitted to the Protein Data Bank at Brookhaven National Laboratory.

29 September 1986; accepted 29 May 1987

The Catalytic Role of the Active Site Aspartic Acid in Serine Proteases

CHARLES S. CRAIK, STEVEN ROCZNIAK,* COREY LARGMAN,† WILLIAM J. RUTTER

The role of the aspartic acid residue in the serine protease catalytic triad Asp, His, and Ser has been tested by replacing Asp¹⁰² of trypsin with Asn by site-directed mutagenesis. The naturally occurring and mutant enzymes were produced in a heterologous expression system, purified to homogeneity, and characterized. At neutral pH the mutant enzyme activity with an ester substrate and with the Ser¹⁹⁵-specific reagent diisopropylfluorophosphate is approximately 10⁴ times less than that of the unmodified enzyme. In contrast to the dramatic loss in reactivity of Ser¹⁹⁵, the mutant trypsin reacts with the His⁵⁷-specific reagent, tosyl-L-lysine chloromethylketone, only five times less efficiently than the unmodified enzyme. Thus, the ability of His⁵⁷ to react with this affinity label is not severely compromised. The catalytic activity of the mutant enzyme increases with increasing pH so that at pH 10.2 the k_{cat} is 6 percent that of trypsin. Kinetic analysis of this novel activity suggests this is due in part to participation of either a titratable base or of hydroxide ion in the catalytic mechanism. By demonstrating the importance of the aspartate residue in catalysis, especially at physiological pH, these experiments provide a rationalization for the evolutionary conservation of the catalytic triad.

SERINE PROTEASES FUNCTION IN many biological systems to hydrolyze specific polypeptide bonds. Trypsin, a well-studied member of this family, catalyzes the hydrolysis of peptide and ester substrates that contain lysyl or arginyl side chains. Serine proteases have the triad of residues Asp¹⁰², His⁵⁷, and Ser¹⁹⁵ at the active site (chymotrypsin numbering system). X-ray crystallographic studies reveal that these three residues are in close proximity, which suggests they may serve as a functional interacting unit responsible for bond formation and cleavage during catalysis (1). Numerous chemical and physical

studies indicate that Ser¹⁹⁵ and His⁵⁷ play crucial roles in catalysis. For example, selective reaction of Ser¹⁹⁵ with diisopropylfluor-

C. S. Craik, Departments of Pharmaceutical Chemistry and of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143-0446. S. Rocznik, C. Largman, W. J. Rutter, Hormone Research Institute and Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143-0448.

*Present address: NutraSweet Company, Mount Prospect, IL 60056.

†Present address: Veterans Administration Hospital, Martinez, CA 94553, and Departments of Internal Medicine and Biological Chemistry, University of California, Davis, CA 95616.

Exhibit 36

A Novel Low-Density Lipoprotein Receptor-Related Protein with Type II Membrane Protein-Like Structure Is Abundant in Heart¹

Yasuhiro Tomita, Dong-Ho Kim, Kenta Magoori, Takahiro Fujino, and Tokuo T. Yamamoto²

Tohoku University Gene Research Center, Sendai 981-8555

Received for publication, May 21, 1998

We report herein the identification of a novel member of the low-density lipoprotein receptor (LDLR) family termed LDLR-related protein 4 (LRP4). Murine LRP4 cDNA encodes a 1113-amino-acid type II membrane-like protein with eight ligand-binding repeats in two clusters. Southern blot analysis of genomic DNA from several different organisms suggests the presence of LRP4 homologues in chicken lacking the gene encoding apolipoprotein E, which is recognized by the ligand-binding repeats of LDLR. LRP4 transcripts were detected almost exclusively in heart in mouse and humans. Despite the presence of the ligand-binding repeats, COS cells transfected with LRP4 did not show surface-binding of β -migrating very-low-density lipoprotein, suggesting that LRP4 plays a role in a pathway other than lipoprotein metabolism.

Key words: LDL receptor family, LDL receptor related protein, membrane protein, receptor.

The low-density lipoprotein receptor (LDLR) family is a growing super gene family that includes LDLR itself (1), apolipoprotein E (apoE) receptor 2 (apoER2) (2, 3), very-low-density lipoprotein receptor (VLDLR) (4, 5), insect vitellogenin receptors (6, 7), LDLR-related protein/ α_2 -macroglobulin receptor (LRP1) (8), a kidney autoantigen gp330/megalin (LRP2) (9, 10), and a recently identified member termed LDLR relative with 11 binding repeats (LR11/sorLA1) (11, 12). All members of this gene family contain the following five structural motifs: (i) complement-type cysteine-rich repeats, termed LDLR ligand-binding repeats or LDLR class A repeats; (ii) cysteine-rich epidermal growth factor (EGF) precursor-type repeats, termed growth factor repeats or LDLR class B repeats; (iii) cysteine-poor spacer regions, with five copies of the sequence YWTD, separating the growth-factor repeats; (iv) a single membrane-spanning region; and (v), a cytoplasmic region with at least one copy of the "NPXY" internalization signal. LDLR is the best characterized protein in this superfamily and the relationship between structure and function for each module of LDLR has been elucidated by analysis of mutations in patients with familial hypercholesterolemia (13, 14).

¹ This work was supported by the Japan Society for the Promotion of Science Grant RFTF97L00803. Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession No. AB013874.

² To whom correspondence should be addressed: Fax: +81-22-263-9295, E-mail: yama@biochem.tohoku.ac.jp

Abbreviations: apoE, apolipoprotein E; apoER2, apolipoprotein E receptor 2; LDLR, low-density lipoprotein receptor; LRP, low-density lipoprotein receptor-related protein; VLDLR, very-low-density lipoprotein receptor; β -VLDL, β -migrating very-low-density lipoprotein.

Among members of the LDLR family, VLDLR and apoER2 most closely resemble LDLR in structure and, like LDLR, bind apoE-rich β -VLDL with high affinity (2-4). In the chicken, VLDLR is expressed almost exclusively in oocytes and mediates uptake of yolk precursors, VLDL and vitellogenin (15). This receptor-mediated process is critical in non-mammalian vertebrate oogenesis: female chicken mutants lacking VLDLR are sterile (16). In contrast to the chicken, mammalian VLDLR mRNA is abundant in heart, skeletal muscle, brain, and adipose tissues (4). Frykman *et al.* have shown that mice lacking VLDLR exhibit modest decreases in body weight, body mass index, and adipose tissue mass, while their plasma cholesterol levels, triacylglycerol levels, and lipoprotein profiles are not altered (17). Furthermore, knockout mice lacking both VLDLR and LDLR exhibit a modest hypercholesterolemia (17), whereas apoE knockout mice exhibit a profound hypercholesterolemia (18). These data suggest the presence of other apoE receptors.

To extend our studies on receptors that may play a role in the clearance of apoE-containing lipoproteins from the circulation, we have been characterizing cDNAs belonging to the LDLR superfamily. In the previous study, we have characterized a new LDLR-related protein termed LRP3 (19). Human and rat LRP3 consist of a 770-amino-acid type I membrane protein with the following regions: a putative signal sequence; two isoleucine/leucine/valine-rich regions with an RGD sequence; two ligand-binding repeat regions; a putative transmembrane region; and a proline-rich cytoplasmic region with a tyrosine-based internalization signal. Despite the presence of the ligand-binding repeats, CHO cells transfected with LRP3 failed to bind β -VLDL.

In this study, we have isolated a near full-length cDNA encoding a new member of the LDLR family, termed

NA
red

Fig. 1

Fig. 1. Nucleotide and deduced amino acid sequence of murine LRP4 cDNA. Nucleotide and amino acid residues are numbered on the left. Nucleotide 1 is the A of the initiator AUG codon. Negative numbers refer to the 5'-untranslated region. Two in-frame translation termination codons at -87 and 3342 are indicated by asterisks. The putative transmembrane region is boxed in black. Cysteine residues are circled and the ligand-binding motif SDE and similar sequences are boxed. Potential N-linked glycosylation sites are underlined and a potential polyadenylation signal is doubly underlined.

LDLR-related protein 4 (LRP4) and describe here the molecular characterization of this new receptor-like protein.

MATERIALS AND METHODS

Standard Procedures—Standard molecular biology techniques were carried out essentially as described by Sambrook et al. (20). Nucleotide sequencing was performed by the dideoxy-chain termination method (21) using M13 primers, T3 and T7, or specific internal primers. Sequence reactions were carried out using Taq DNA polymerase with fluorescently labeled nucleotides on an Applied Biosystems Model 373A DNA sequencer. To analyze RNA in murine and human tissues, commercially available Northern blots (Clontech) were used for Northern blot analysis.

cDNA Cloning—A murine heart cDNA library was constructed in pBluescript vector using poly(A) RNA and the cDNA synthesis kit from Pharmacia. The library was screened with a mixture of degenerative oligonucleotides corresponding to a highly conserved amino acid sequence, WRCDGD, among the ligand-binding domains of LDLR, VLDLR, and apoER2: 5'-TGG(A/C)G(A/C/G/T)TG(C/T)-GA(C/T)GG(A/C/G/T)GA-3'. Positive clones hybridizing

with the oligonucleotide probe were the reprobated with LDLR and VLDLR probes to eliminate cDNAs for these receptors. By screening 5×10^5 clones, we obtained one positive clone that hybridized with the oligonucleotide probe alone.

"Zoo" Southern Blot Analysis—Genomic DNAs (10 μ g) prepared from a normal man, a male BALB/c mouse, a white Leghorn hen, and a female *Xenopus laevis* were digested with a large excess of *Eco*RI for electrophoresis in a 0.8% agarose gel, then transferred onto a nylon membrane. The membrane was hybridized with the entire region of murine LRP4 cDNA. Hybridization was at 42°C in $5 \times$ SSC, $5 \times$ Denhardt's solution, 200 μ g/ml denatured salmon sperm DNA, 50% (v/v) formamide, and 1% (w/v) SDS. The blot was then washed twice with $0.3 \times$ SSC and 1% (w/v) SDS at 60°C, followed by autoradiography.

Expression of LRP4 cDNA in COS-7 Cells—To construct an LRP4 expression plasmid (pLRP4-SR α), the entire coding region of murine LRP4 cDNA was inserted into an expression vector (pCDL-SR α 296) (22) by multiple ligations of restriction fragments. The expression plasmid was transfected into COS-7 cells according to the transfection protocol described by Chen and Okayama (23).

Lipoprotein Binding Assay—Rabbit β -VLDL (1.006 g/ml) was prepared from the plasma of 1% cholesterol-fed animals (24). 125 I-labeled β -VLDL was prepared (25) and its binding by the transfected cells was assayed according to the procedure described previously (2).

RESULTS

Isolation and Characterization of Murine LRP4 cDNA—A near full-length cDNA encoding a new member of the LDLR family, designated LDLR-related protein 4 (LRP4),

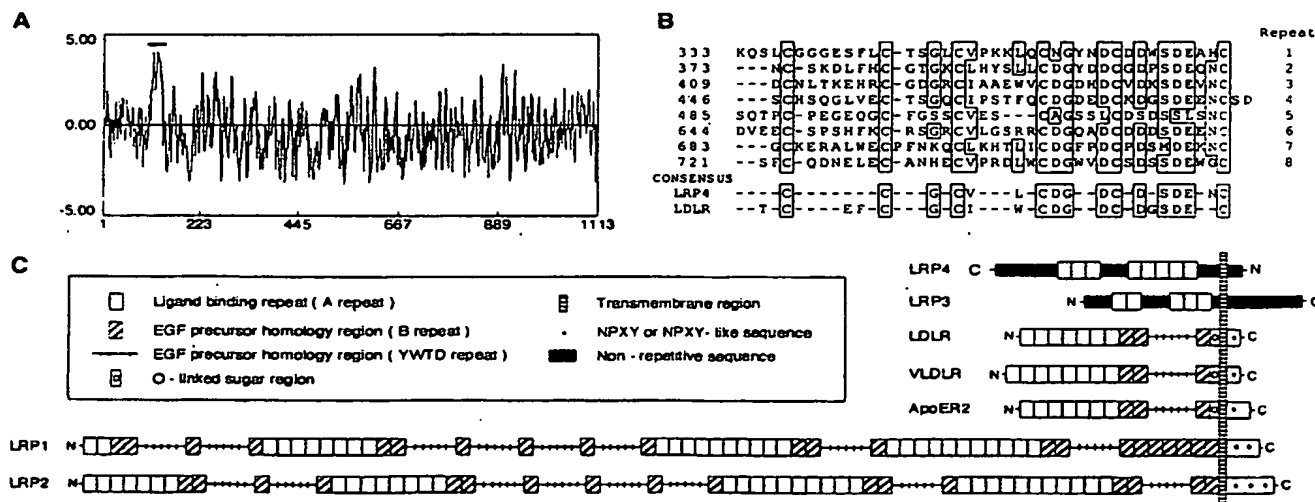


Fig. 2. Functional regions in LRP4. (A) Hydropathy plot analysis of the murine LRP4 protein. The numbers on the x-axis correspond to the positions of the amino acid residues in the protein. The putative transmembrane (TM) region is shown by a thick line. (B) Comparison of the amino acids in the eight ligand-binding repeats of murine LRP4. Amino acid alignment was optimized and gaps were introduced to

match the six cysteine residues in each repeat. Amino acid residues conserved in more than 50% of the repeats are boxed and shown below as a consensus sequence. The consensus sequence of the ligand-binding repeats of human LDLR (1) is also represented. (C) Schematic representation of LRP4-4, apoER2, LDLR, and VLDLR.

Low-Density Lipoprotein Receptor-Related Protein 4

was isolated from a murine heart cDNA library by using a mixture of degenerative oligonucleotides corresponding to the highly conserved amino acid sequence WRCDGD among the ligand-binding domains of LDLR, VLDLR, and apoER2. Figure 1 shows the nucleotide and deduced amino acids sequences of the cDNA, which has an open reading frame of 3,339 bp corresponding of 1,113 amino acids with a calculated molecular mass of approximately 123 kDa. The putative initial methionine was preceded by an in-frame termination codon present 87 nucleotides upstream.

A hydropathy plot (26) of the deduced amino acid sequence of murine LRP4 shows the presence of a hydrophobic region at amino acid residues 113-133 (boxed in black in Fig. 1 and identified with thick lines in Fig. 2A). This hydrophobic sequence of 21 amino acids strongly resembles the transmembrane region of membrane proteins, being flanked by a positively charged amino acid (arginine) on the N-terminal side. This structural feature suggests that LRP4 has a type II transmembrane protein structure (amino terminus in the cytosol).

The C-terminal side of the putative transmembrane domain contains two clusters of cysteine-rich repeats that resemble the ligand binding repeats (class A motifs) of LDLR: one cluster contains three repeats and the other has five (Fig. 2, B and C). Each repeat has six completely conserved cysteines and a highly conserved C-terminal SDE tripeptide, which forms a part of the ligand-binding site of LDLR (Fig. 2B). Unlike LDLR, VLDLR, apoER2, LRP1, and LRP2, there are neither YWTD repeats nor growth factor repeats (class B motifs) in the murine LRP4 sequence (Fig. 2C).

The cytoplasmic domains of LDLR, VLDLR, apoER2, LRP1, and LRP2 contain one or two copies of a highly conserved coated pit signal, FXNPXY (23). In the putative cytoplasmic region (N-terminus), we found neither a typical FXNPXY sequence nor a similar tyrosine-based sequence (27). Further studies are required to determine whether LRP4 may function as an endocytic receptor.

Southern Blot Analysis of the LRP4 Genes in Various Species—To test the possibility that LRP4 homologue genes might also be present in nonmammalian vertebrates (known to lack the apoE gene), Southern blot analysis of genomic DNA from several different organisms was carried out. This "zoo blot" (containing DNAs of humans, mouse, chicken, and frog) was hybridized with the entire coding region of the murine cDNA under relatively stringent conditions (see "MATERIALS AND METHODS"). As shown in Fig. 3, intense hybridization signals are present in mouse,

and fainter but significant signals can also be detected in human and chicken DNAs. These data suggest the presence of LRP4 homologues in chicken lacking the gene encoding apoE, which is recognized by the ligand-binding repeats of mammalian LDLR, VLDLR, and apoER2.

Expression of LRP4 Transcripts—Northern blot analysis of RNA from various murine tissues revealed hybridization of the LRP4 probe to a major transcript of 5.0 kb in mouse, with the highest expression in heart, relatively high levels in testis, and much lower levels in kidney and lung (Fig. 4A). Figure 4B shows a blot hybridization of RNA from various human tissues probed with the murine cDNA. In human tissues, major transcripts of 5, 2.6, and 2.3 kb and a minor transcript of 4 kb are detected almost exclusively in heart. A fainter but significant signal of 2 kb can also be detected in skeletal muscle and testis. The transcripts of 2.0, 2.3, 2.6, and 4 kb detected in human tissues may be a consequence of alternative splicing.

β -VLDL Binding—To test the possibility that LRP4 might bind apoE-rich β -VLDL (as do LDLR, VLDLR, and apoER2), an expression plasmid containing the entire coding region of murine LRP4 cDNA was constructed and introduced into COS-7 cells, and ligand-binding activity was measured using 125 I-labeled β -VLDL. As shown in Fig.

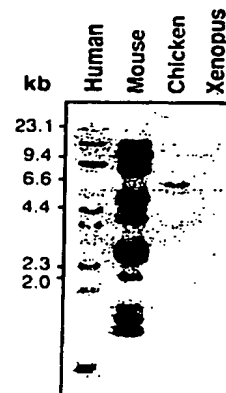


Fig. 3. Genomic Southern blot analysis of LRP4-related sequences in various eukaryotic species. A blot containing 10 μ g of *Eco*RI-digested DNA from the indicated species was hybridized with the entire coding region of murine LRP4 cDNA under the conditions described in "MATERIALS AND METHODS" and exposed to Kodak XAR-5 film with an intensifying screen at -80°C for 16 h.

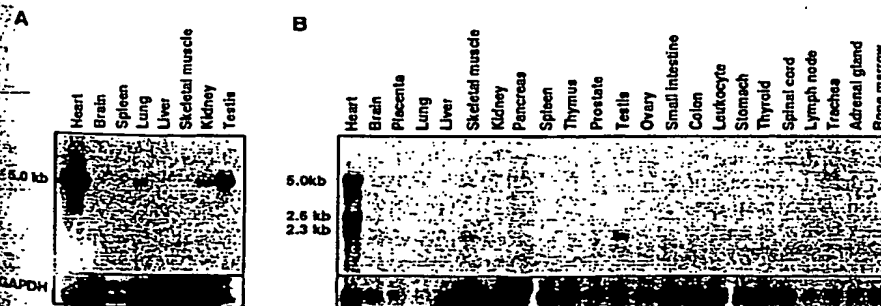
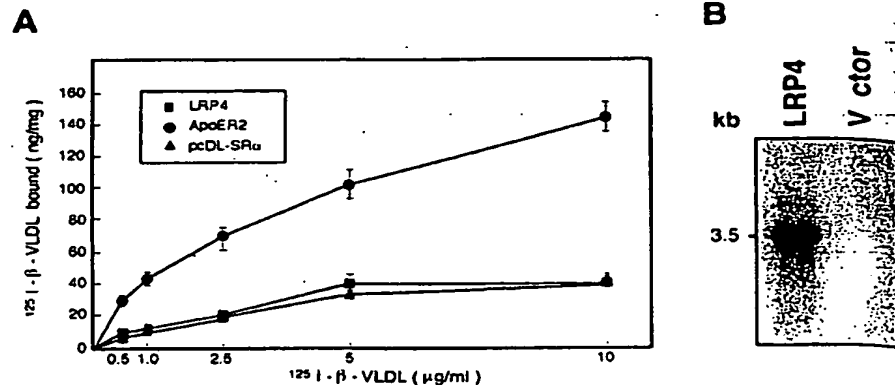


Fig. 4. Expression of LRP4 transcripts in mouse (A) and humans (B). Poly(A) RNA (2 μ g) from the indicated murine (A) and human (B) tissues was probed with ^{32}P -labeled murine LRP4 cDNA. The filters were exposed to Kodak XAR-5 film with an intensifying screen at -80°C for 14 h. Control hybridization with a rat glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is shown in the lower portion.

Fig. 5. Transient expression of LRP4 in COS cells. (A) Surface binding of 125 I-labeled β -VLDL. COS cells transfected with an expression plasmid encoding murine LRP4 (pLRP4-SR α), human apoER2 (pNR1), or the parental vector of pLRP4-SR α (pcDL-SR α 296) were incubated for 2 h at 4°C with the indicated concentrations of 125 I- β -VLDL (540 cpm/ng), after which the values for surface-bound β -VLDL were determined as described under "MATERIALS AND METHODS." (B) Northern blot analysis of LRP4 transcripts in COS cells transfected with murine LRP4 expression plasmid (LRP4), or the parental vector (pcDL-SR α 296). Total RNA (10 μ g) from the indicated transfected cells was probed with 32 P-labeled murine LRP4 cDNA. The filter was exposed to Kodak XAR-5 film with an intensifying screen at -80°C for 12 h.



5A, the level of surface bound β -VLDL in LRP4-transfected cells was similar to those in cells transfected with equal amounts of the parental vector, despite the high levels of accumulation of 3.0-kb LRP4 mRNA (lacking approximately 2.0 kb in the 3'-untranslated region) in the LRP4-transfected cells (Fig. 5B). In control experiments, marked induction of 125 I- β -VLDL binding was observed in cells transfected with human apoER2.

DISCUSSION

In the present study, we have shown the structure and expression of a novel member of the LDLR family termed LRP4. The most interesting feature of LRP4 is that, unlike other members of the LDLR family, this protein has a type II membrane protein-like structure. The hydropathy plot analysis shows the presence of a hydrophobic region at amino acid residues 113-133 of murine LRP4. There are eight ligand-binding repeats clustered into two regions in the C-terminal side of this putative transmembrane region. Based on the presence of ligand-binding repeats in the extracellular regions of other LDLR family members, it seems reasonable to predict that the C-terminal side of the putative transmembrane region constitutes the extracellular region of the protein.

Despite the presence of eight ligand-binding repeats, COS cells transfected with LRP4 failed to bind β -VLDL, suggesting that LRP4 does not function in lipoprotein metabolism. Of the four clusters of ligand-binding repeats in LRP2, the recognition site for apoE has been mapped to the second cluster (28). This suggests that these clusters are not functionally equal, despite their structural similarity. Therefore, the ligand-binding repeats in LRP4 may be functionally different from those in other family members that bind β -VLDL.

Although the exact function and ligands of LRP4 remain unclear, the abundant expression of LRP4 transcripts in heart is noteworthy. Based on the structural features of LRP4 and its almost exclusive expression in the heart, LRP4 may play a role as a surface receptor that is related to cardiac function. Further studies are necessary to elucidate the exact role of this structurally interesting molecule.

We thank Kyoko Ogamo and Nami Suzuki for secretarial assistance.

REFERENCES

1. Yamamoto, T., Davis, C.G., Brown, M.S., Schneider, W.J., Casey, M.L., Goldstein, J.L., and Russell, D.W. (1984) The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. *Cell* 39, 27-38
2. Kim, D.H., Iijima, H., Goto, K., Sakai, J., Ishii, H., Kim, H.J., Suzuki, H., Kondo, H., Saeki, S., and Yamamoto, T. (1996) Human apolipoprotein E receptor 2. A novel lipoprotein receptor of the low density lipoprotein receptor family predominantly expressed in brain. *J. Biol. Chem.* 271, 8373-8380
3. Kim, D.H., Magoori, K., Inoue, T.R., Mao, C.C., Kim, H.J., Suzuki, H., Fujita, T., Endo, Y., Saeki, S., and Yamamoto, T.T. (1997) Exon/intron organization, chromosome localization, alternative splicing, and transcription units of the human apolipoprotein E receptor 2 gene. *J. Biol. Chem.* 272, 8498-8504
4. Takahashi, S., Kawarabayashi, Y., Nakai, T., Sakai, J., and Yamamoto, T. (1992) Rabbit very low density lipoprotein receptor: a low density lipoprotein receptor-like protein with distinct ligand specificity. *Proc. Natl. Acad. Sci. USA* 89, 9252-9256
5. Sakai, J., Hoshino, A., Takahashi, S., Miura, Y., Ishii, H., Suzuki, H., Kawarabayashi, Y., and Yamamoto, T. (1994) Structure, chromosome location, and expression of the human very low density lipoprotein receptor gene. *J. Biol. Chem.* 269, 2173-2182
6. Schonbaum, C.P., Lee, S., and Mahowald, A.P. (1995) The *Drosophila* yolkless gene encodes a vitellogenin receptor belonging to the low density lipoprotein receptor superfamily. *Proc. Natl. Acad. Sci. USA* 92, 1485-1489
7. Sappington, T.W., Kokoza, V.A., Cho, W.L., and Raikhel, A.S. (1996) Molecular characterization of the mosquito vitellogenin receptor reveals unexpected high homology to the *Drosophila* yolk protein receptor. *Proc. Natl. Acad. Sci. USA* 93, 8934-8939
8. Herz, J., Hamann, U., Røge, S., Myklebost, O., Gausepohl, H., and Stanley, K.K. (1988) Surface location and high affinity for calcium of a 500 kD liver membrane protein closely related to the LDL-receptor suggest a physiological role as lipoprotein receptor. *EMBO J.* 7, 4119-4127
9. Raychowdhury, R., Niles, J.L., McCluskey, R.T., and Smith, J.A. (1989) Autoimmune target in Heymann nephritis is a glycoprotein with homology to the LDL receptor. *Science* 244, 1163-1165
10. Saito, A., Pietromonaco, S., Loo, A.K., and Farquhar, M.G. (1994) Complete cloning and sequencing of rat gp330/"megalin," a distinctive member of the low density lipoprotein receptor gene family. *Proc. Natl. Acad. Sci. USA* 91, 9725-9729

11. Bujo, H., Hermann, M., Schneider, W.J., and Nimpf, J. (1996) A new branch of the LDL-receptor family tree: VLDL-receptors. *Z. Gastroenterol.* 3, 124-126
12. Jacobsen, L., Madsen, P., Moestrup, S.K., Lund, A.H., Tommerup, N., Nykjaer, A., Sottrup-Jensen, L., Gliemann, J., and Petersen, C.M. (1996) Molecular characterization of a novel human hybrid-type receptor that binds the alpha₂-macroglobulin receptor-associated protein. *J. Biol. Chem.* 271, 31379-31383
13. Russell, D.W., Lehman, M.A., Südhof, T.C., Yamamoto, T., Davis, C.G., Hobbs, H.H., Brown, M.S., and Goldstein, J.L. (1987) The LDL receptor in familial hypercholesterolemia: Use of human mutations to dissect a membrane protein. *Cold Spring Harbor Symp. Quant. Biol.* 51, 811-819
14. Hobbs, H.H., Russell, D.W., Brown, M.S., and Goldstein, J.L. (1990) The LDL receptor locus in familial hypercholesterolemia: mutational analysis of a membrane protein. *Annu. Rev. Genet.* 24, 133-170
15. Bujo, H., Hermann, M., Kaderli, M.O., Jacobsen, L., Sugawara, S., Nimpf, J., Yamamoto, T., and Schneider, W.J. (1994) Chicken oocyte growth is mediated by an eight ligand binding repeat member of the LDL receptor family. *EMBO J.* 13, 5165-5175
16. Bujo, H., Yamamoto, T., Hayashi, K., Hermann, M., Nimpf, J., and Schneider, W.J. (1995) Mutant oocyte low density lipoprotein receptor gene family member causes atherosclerosis and female sterility. *Proc. Natl. Acad. Sci. USA* 92, 9905-9909
17. Frykman, P.K., Brown, M.S., Yamamoto, T., Goldstein, J.L., and Herz, J. (1995) Normal plasma lipoproteins and fertility in gene-targeted mice homozygous for a disruption in the gene encoding very low density lipoprotein receptor. *Proc. Natl. Acad. Sci. USA* 92, 8453-8457
18. Ishibashi, S., Herz, J., Maeda, N., Goldstein, J.L., and Brown, M.S. (1994) The two-receptor model of lipoprotein clearance: test of the hypothesis in "knockout" mice lacking the low density lipoprotein receptor, apolipoprotein E, or both proteins. *Proc. Natl. Acad. Sci. USA* 91, 4431-4435
19. Ishii, H., Kim, D.H., Fujita, T., Endo, Y., Saeki, S., and Yamamoto, T.T. (1998) cDNA cloning of a new low density lipoprotein receptor-related protein and mapping of its gene (LRP3) to chromosome bands 19q12-13.2. *Genomics* 51, 132-135
20. Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
21. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463-5467
22. Takebe, Y., Seiki, M., Fujisawa, J., Hoy, P., Yokota, K., Arai, K., Yoshida, M., and Arai, N. (1988) SR alpha promoter: an efficient and versatile mammalian cDNA expression system composed of the simian virus 40 early promoter and the R-U5 segment of human T-cell leukemia virus type 1 long terminal repeat. *Mol. Cell. Biol.* 8, 466-472
23. Chen, C.A. and Okayama, H. (1987) High-efficiency transformation of mammalian cells by plasmid DNA. *Mol. Cell. Biol.* 7, 2745-2752
24. Kovanen, P.T., Brown, M.S., Basu, S.K., Bilheimer, D.W., and Goldstein, J.L. (1981) Saturation and suppression of hepatic lipoprotein receptors: a mechanism for the hypercholesterolemia of cholesterol-fed rabbits. *Proc. Natl. Acad. Sci. USA* 78, 1396-1400
25. Goldstein, J.L., Basu, S.K., and Brown, M.S. (1983) Receptor-mediated endocytosis of low-density lipoprotein in cultured cells. *Methods Enzymol.* 98, 241-260
26. Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105-132
27. Naim, H.Y. and Roth, M.G. (1994) Characteristics of the internalization signal in the Y543 influenza virus hemagglutinin suggest a model for recognition of internalization signals containing tyrosine. *J. Biol. Chem.* 269, 3928-3933
28. Orlando, R.A., Exner, M., Czekay, R.P., Yamazaki, H., Saito, A., Ullrich, R., Kerjaschki, D., and Farquhar, M.G. (1997) Identification of the second cluster of ligand-binding repeats in megalin as a site for receptor-ligand interactions. *Proc. Natl. Acad. Sci. USA* 94, 2368-2373



Exhibit 37

Hepsin, a Cell Membrane-associated Protease

CHARACTERIZATION, TISSUE DISTRIBUTION, AND GENE LOCALIZATION*

(Received for publication, August 2, 1990)

Akihiko Tsuji, Adrian Torres-Rosado, Toshiro Arai, Michelle M. Le Beau[§], Richard S. Lemons[¶],
Shan-Ho Chou^{**}, and Kotoku Kurachi^{††}

From the Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, the [†]Section of Hematology and Oncology, University of Chicago, Chicago, Illinois 60637, the [¶]Department of Pediatrics, University of Utah, Salt Lake City, Utah 84132, and the [§]Department of Biochemistry, University of Washington, Seattle, Washington 98109

Hepsin, a putative membrane-bound serine protease, was originally identified as a human liver cDNA clone (Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* 27, 1067-1074). In the present study the human hepsin gene was localized to chromosome 19 at q11-13.2. The messenger RNA of hepsin is 1.85 kilobases in size and present in most tissues, with the highest level in liver. Hepsin is synthesized as a single polypeptide chain, and its mature form of 51 kDa was found in various mammalian cells including HepG2 cells and baby hamster kidney cells. It is present in the plasma-membrane in a molecular orientation of type II membrane-associated proteins, with its catalytic subunit (carboxyl-terminal half) at the cell surface, and its amino terminus facing the cytosol. Hepsin is found neither in cytosol nor in culture media. The results obtained suggest that hepsin has an important role(s) in cell growth and function.

Proteases play important roles in a number of physiological and pathological processes such as protein catabolism, blood coagulation, fibrinolysis, and in the complement system (1-3). The importance of proteases in many phenomena including cell proliferation, inflammation, development, tumor growth, and metastasis are also well described. Their involvement in carcinogenesis as well as in cell growth is further supported by the anticarcinogenic and anti-cell growth effects of protease inhibitors (4, 5). Most of these are non-membrane bound intra- or extracellular proteases. Recently, several membrane-associated proteases have been described. A cell surface protease with molecular weight of 67,000 has been reported (5-7). This protease, which is inhibited by α_1 -antitrypsin (5), was found to be essential for cell proliferation and was suggested to be involved in various biological processes of cells, in addition to the degradation of extracellular matrix proteins. Guanidinobenzoate, which can cleave fibronectin at Gly-Arg-Gly-Asp, the sequence involved in the attachment of fibronectin to cell surfaces, has been described (8-10). This protease is located on the surface of most tumor cells, as well as in the fluid surrounding tumor cells. A fluorescent compet-

itive inhibitor has also been used to localize this protease on the tumor cell surface (9). A trypsin-like membrane-associated protease of an estimated molecular weight of 120,000 which is present in the liver has been proposed to be involved in membrane protein turnover (11). A membrane-bound trypsin-like protease has also been recognized in other cells such as neuroblastoma cells (12). More recently, a 170-kDa membrane-bound protease (gelatinase) has been implicated in melanoma cell invasiveness (13). As described in these reports the cell surface proteases are considered to play an important role(s) in cell growth, cell invasion of other tissues (such as in metastasis), angiogenesis, and tissue rearrangement, in addition to various other cellular processes.

Hepsin is a putative serine protease of 417 amino acid residues originally identified from cDNA clones isolated from human liver cDNA libraries (14). In a previous study, a synthetic oligonucleotide probe for the amino acid sequence Met-Phe-Cys-Ala-Gly, which is common to many serine proteases, was successfully employed to isolate a number of known and novel proteases including hepsin. Hepsin contains a short hydrophobic amino acid sequence in the region near the amino terminus while its carboxyl-terminal half is a typical serine protease module. The hydrophobic sequence, composed of 27 amino acid residues, is very similar to the typical lipid bilayer membrane-spanning sequences found in many other membrane-associated proteins (14). In our preliminary immunostaining study, hepsin was found to be present in cultured cells such as HepG2 and baby hamster kidney (BHK)¹ cells (15). It is highly likely that hepsin may have a role(s) similar to other cell membrane-bound proteases described above in cell growth and in other cell functions. Presently, however, the protein chemical and enzymatic properties as well as the precise biological role(s) of hepsin are not known.

In this report, we describe evidence that demonstrates the actual existence of hepsin in cells. This includes determination of the estimated molecular weight of cellular hepsin, its subcellular localization, topology at the cell surface, chromosomal localization of its gene, as well as its tissue distribution of expression.

EXPERIMENTAL PROCEDURES

Materials—Keyhole limpet hemocyanin and bovine pancreatic trypsin were obtained from Sigma. Freund's adjuvant was purchased from Difco. Synthetic peptides were made by an automated peptide synthesizer (Applied Biosystems, model 438) employing solid-phase t-butoxycarboxyl chemistry. These peptides had free α -carboxyl

* This work was supported in part by National Institutes of Health Grant HL38644 (to K. K.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[§] Scholar of the Leukemia Society of America.

[¶] Associate member of the Howard Hughes Medical Institute, University of Washington.

^{††} To whom all correspondence and reprint requests should be addressed.

¹ The abbreviations used are: BHK, baby hamster kidney; PBS, phosphate-buffered saline; SDS, sodium dodecyl sulfate; EGTA, [ethylenedis(oxyethylenenitrilo)]tetraacetic acid; kb, kilobase.

groups. Activated CH-Sepharose 4B and Percoll were obtained from Pharmacia. Tissue culture supplies and proteinase K were purchased from Gibco/BRL (Life Technologies, Inc.). ^{14}C -Labeled size marker protein kits were obtained from Du Pont-New England Nuclear. All radioactive nucleotides were purchased from Amersham Corp. The protein assay kit as well as peroxidase-conjugated goat anti-rabbit IgG were obtained from Bio-Rad. Adenosine 5'-phosphate and 4-chloro-1-naphthol were purchased from Sigma. Nylon membranes (GeneScreen Plus®) and the reticulocyte cell-free translation kit were from New England BioLab (Du Pont).

Preparation of Antibodies—Five synthetic peptides (P1, amino acid 1–17; P2, 246–257; P3, 294–305; P4, 360–372; and P5, 398–417) corresponding to the amino acid sequence of hepsin predicted from the cDNA sequence (14) were employed to raise antibodies. P1, PM (equimolar mixture of P2, P3, P4), and P5 correspond to the sequences of the amino-terminal region, the catalytic subunit, and the carboxyl-terminal region, respectively. P1, PM, or P5 were separately coupled to the keyhole limpet hemocyanin by using glutaraldehyde as a coupling agent as described by Reichlin (16). Rabbits were immunized with a mixture of keyhole limpet hemocyanin-peptide conjugate with Freund's adjuvant as follows: 5 mg of the conjugate in complete Freund's adjuvant was injected subcutaneously on day 1, and 1 mg of conjugate in incomplete Freund's adjuvant (1:1) was injected on days 14, 21, and 28. After the third and fourth injection on days 14 and 28, animals were bled from the ear vein to test the titer. After the fifth week, blood samples were collected from the animals by heart puncture, and were then used to prepare affinity purified antibodies.

Affinity purification of these antibodies was carried out as follows: peptide column was prepared by adding peptides (10 mg dissolved in 20 ml of 0.1 M NaHCO_3 , pH 9.0) to the activated CH-Sepharose 4B (1 g dry weight) (Pharmacia) according to the manufacturer's instructions. Antiserum (3 ml), which was incubated with 8 mg of hemocyanin for 1 h at room temperature, was applied to the column (2.6 ml) followed by extensive washing with 10 mM sodium phosphate, pH 7.4, containing 0.15 M NaCl (PBS). The bound immunoglobulins were then eluted with 0.1 M glycine-HCl buffer, pH 2.3, into 0.2 ml of 1 M Tris-HCl buffer, pH 7.0. The eluate was dialyzed against PBS and stored at -80°C until use. Affinity purified antibodies prepared against peptides P1, PM, and P5 were designated HAbP1, HAbPM, and HAbP5, respectively. Immunoblot tests showed that HAbPM and HAbP5 were highly specific, while HAbP1 was not, probably due to cross-reactivity with similar amino acid sequences apparently present in other proteins.

Cell Culture—HepG2 cells and BHK cells were cultured in Eagle's minimum essential medium (Gibco) supplemented with streptomycin, penicillin, and 10% fetal calf serum in a 5% CO_2 incubator at 37°C .

Fractionation of Cellular Components by Percoll Density Gradient Centrifugation—HepG2 cells ($\sim 6 \times 10^6$ cells) were harvested by scraping, washed twice with PBS (1000 rpm for 5 min at 4°C), and resuspended in 3 ml of ice-cold STE solution (0.25 M sucrose, 10 mM Tris-HCl buffer, pH 7.5, containing 2 mM EGTA) followed by homogenization with a Tekmar Ultra-Turrax tissue homogenizer for 15 s. Plasma membrane and mitochondrial fractions were isolated by the method of Belsham *et al.* (17) with minor modifications. Briefly, the homogenates were centrifuged at $100 \times g$ for 1 min. The pellets obtained were resuspended in 2 ml of STE solution, homogenized, and centrifuged. The two supernatants were combined and centrifuged at $5000 \times g$ for 15 min. A fraction (0.5 ml) of the pellet was suspended in 1.0 ml of STE solution, dispersed in 10 ml of iso-osmotic Percoll solution (7 volumes of Percoll, 1 volume of 2 M sucrose, 80 mM Tris-HCl buffer, pH 7.5, containing 8 mM EGTA and 32 volumes of STE solution), and centrifuged for 20 min at $10,000 \times g$ (Sorvall and RC5C with SS34 rotor). Two membrane bands, one immediately below the surface (plasma membrane) and the other close to the bottom (mitochondria) were separately collected into 4 volumes of 10 mM Tris-HCl buffer, pH 7.5, containing 0.15 M NaCl. The two fractions collected were then centrifuged at $10,000 \times g$ for 3 min to obtain membrane samples. The enrichment of the plasma membrane prepared was monitored by assaying a plasma membrane-associated lipoprotein, 5'-nucleotidase, according to Windell and Unkeless (18). The purity of the membrane preparation was further tested by assaying activities of glucose 6-phosphatase (microsome marker) and succinate-cytochrome *c* reductase (mitochondria marker) according to Sottocasa *et al.* (19) with minor modifications. The microsome fraction used as a control in the assay was prepared as previously described (19, 20).

An aliquot of the cell homogenates (above) was subjected to cen-

trifugation at $100,000 \times g$ for 30 min at 4°C in a SW41.1 rotor (Beckman model L5–50 centrifuge). The supernatant collected was used as the cytosol fraction. The nuclear fraction was prepared from cell homogenates by sucrose density gradient centrifugation according to Blobel and Potter (21).

Plasma membrane, mitochondria, and nuclear fractions were solubilized with 0.2 ml of 10 mM Tris-HCl buffer, pH 7.5, containing 0.15 M NaCl and 0.5% (w/v) Nonidet P-40 and used for immunoblot analysis.

Immunoblot Analysis—Protein concentration of the samples was determined by the method of Bradford (22) with minor modifications. Proteins of solubilized plasma membranes, mitochondria, nuclei, cytosol, as well as culture media, were adjusted to a concentration of 0.5 mg/ml with gel loading buffer (62.5 mM Tris-HCl, pH 6.8, containing 10% glycerol, and 2% SDS) and incubated at 4°C for 12 h or at 95°C for 3 min. An aliquot (7.5 μg of proteins) of the sample was subjected to SDS-polyacrylamide gel (12%) electrophoresis employing a Bio-Rad mini gel apparatus. The electrophoresed proteins were transferred to a nitrocellulose filter according to Towbin *et al.* (23). The blotted filter was blocked with 3% bovine serum albumin in 50 mM Tris-HCl, pH 7.5, containing 0.15 M NaCl (TBS) at 37°C for 30 min, followed by incubation at room temperature for 2 h with antibodies (P5) raised against the synthetic peptide containing the carboxyl-terminal sequence of hepsin (500-fold dilution in TBS containing 1% bovine serum albumin). The filter was washed 3 times with TBS containing 0.05% Tween 20 and incubated at room temperature for 2 h with horseradish peroxidase-conjugated goat anti-rabbit IgG which was diluted 1000-fold. The filters were then incubated with TBS containing 4-chloro-1-naphthol (0.5 mg/ml) for 30 min at room temperature.

Proteolysis of HepG2 Cells—Mild proteolysis of HepG2 cells to test the topology of hepsin at the cell surface was carried out as follows: HepG2 cells (about 90% confluency) in nine 10-cm culture dishes (total of about 4.5×10^7 cells) were washed twice with phosphate-buffered saline (0.15 M NaCl, 8 mM Na_2HPO_4 , 0.6 mM KH_2PO_4), pH 7.4, and incubated in the buffer for 30 min on ice with or without 10 $\mu\text{g}/\text{ml}$ proteinase K or 100 $\mu\text{g}/\text{ml}$ bovine pancreatic trypsin. Under these conditions, HepG2 cells did not significantly lose their viability. Cells were then washed twice with the phosphate buffer and used for preparing plasma membrane proteins as described above. Aliquots (20 μg each) of protein samples were subjected to immunoblot analysis as described above employing the affinity-purified antibody, P5.

Fluorescent Immunostaining of Cultured Cells—Cells were maintained at 37°C in 5% CO_2 in minimum essential medium containing 10% fetal calf serum and antibiotics. Cells grown to subconfluency on coverslips (8 wells/slide; Miles Laboratories) were fixed at room temperature for 10 min with 2% paraformaldehyde and 0.2% glutaraldehyde in PBS containing Ca^{2+} and Mg^{2+} (Gibco). After rinsing several times with PBS, cells were incubated with goat serum at a dilution of 1:20 in PBS at room temperature for 15 min to block nonspecific binding of the antibody. After several additional rinses with PBS, cells were incubated with purified antisynthetic peptide IgG (2–5 $\mu\text{g}/\text{ml}$ of PM which recognizes the middle portion of the putative catalytic subunit) in PBS containing bovine serum albumin (1 mg/ml) with and without 0.05% Triton X-100 for 2 h in humidified Petri dishes. The bound IgG was visualized by incubating for 30 min with goat anti-rabbit IgG labeled with fluorescein isothiocyanate (diluted 1:50 with PBS). In control experiments: 1) the antibodies were preincubated with synthetic peptides (1 mg/ml of PM) used for raising antibodies before incubating with cells; or 2) PBS containing no antibodies with or without synthetic peptides (1 mg/ml) was added to cells; or 3) anti-hepsin antibodies were replaced with anti-human blood coagulation factor IX. For testing any intracellular immunostaining, cells were treated with 0.5% Triton X-100 for 3–5 min before incubating with the antibodies (HAbPM). The cells were immediately examined by fluorescence microscopy and photographed. In this experiment, HAbP1 antibodies (specific for the amino-terminal region) were not employed because their specificity was found to be low in immunoblot analysis and they recognized not only the 51-kDa band but also a significant number of other bands.

RNA Blot Analysis—Total RNAs of various baboon tissues were prepared by the guanidinium isothiocyanate method described by Chomczynski and Sacchi (24). RNA preparations (20 μg for each tissue) were electrophoresed in a 1.5% agarose gel containing 6.7% formaldehyde in 20 mM phosphate buffer, pH 7.0 (25). The agarose gels were then blotted onto GeneScreen Plus® membranes (Du Pont/New England Nuclear), followed by baking for 2 h at 80°C . A hepsin cDNA (1.8 kb) (14) was labeled with [α - ^{32}P]dCTP by using an

oligolabeling kit (Pharmacia) to a specific activity of about 1×10^9 cpm/ μ g. Prehybridization, hybridization with the radiolabeled cDNA probe, and washing were carried out as described by the manufacturer for the GeneScreen Plus[®] membrane. The membrane was then exposed to x-ray film (Kodak X-Omat AR) at -70°C . A ribosomal RNA gene probe was used to confirm the presence of RNAs in each lane of the blot.

Molecular Mapping of the Gene Locus—A panel of somatic cell hybrids for mapping was established by PEG 1000-mediated cell fusion of human VA2, A549, IMR90 fibroblast or peripheral human lymphocyte cells to either Chinese E36 or Syrian BHK-B1 hamster cells as previously described (26). A panel of hybrids for mapping was established after characterization for their human chromosome content by screening up to 34 gene enzyme systems and, in selected cases, by cytogenetic analyses. ³²P-Labeled hepsin cDNA (1.8 kb) ($1-3 \times 10^9$ dpm/ μ g) was hybridized to DNA blots of these hybrids and controls which had been digested to completion with *Hind*III, *Bam*HI, or *Eco*RI, electrophoresed, and blotted as described (26).

In situ chromosomal hybridization was carried out as follows: human metaphase cells were prepared from phytohemagglutinin-stimulated peripheral blood lymphocytes (27). A radiolabeled, hepsin-specific cDNA probe was prepared by nick translation of the entire plasmid with all four deoxynucleoside triphosphates ³H-labeled to a specific activity of $1-2 \times 10^8$ dpm/ μ g. *In situ* hybridization was performed as described previously (27). Metaphase cells were hybridized at 2.0 and 4.0 ng of probe/ml of hybridization mixture. Autoradiographs were exposed for 11 days.

Cell-free Transcription of Hepsin cDNA and in Vitro Translation—Hepsin cDNA (1.8 kb) (14) was inserted into the pSG5 vector (Stratagene) for both orientations at the unique *Eco*RI site downstream of the T7 promoter. The chimeric plasmid was then transfected into *Escherichia coli* TB-1 cells and amplified followed by preparation employing the alkaline-SDS method and CsCl gradient ultracentrifugation. The plasmids were linearized by digestion with *Xba*I located downstream of the insert in the vector sequence, followed by incubation with proteinase K (50 μ g/ml) at 37°C for 30 min. The reaction mixture was extracted twice with phenol/chloroform (1:1) and ethanol precipitated prior to subjecting it to transcription reactions. The linearized plasmid DNAs were dissolved in TE buffer (10 mM Tris-HCl, pH 7.4, 0.1 mM EDTA prepared with diethyl-pyrocyanate-treated water) and employed as a template for transcription reactions. Cell-free transcription was carried out at 37°C for 30 min with T7 RNA polymerase using an mRNA capping kit (Stratagene) according to the manufacturer's instructions. The transcription reaction mixture was then added to 25 units of RNase free-DNase I followed by an additional incubation for 5 min at 37°C . Synthesized RNA was precipitated with ethanol after extracting once with phenol/chloroform (1:1), dissolved in TE buffer, and employed in translation reactions. The RNAs synthesized were quantitated by reading the absorbance at 260 nm. The size of the RNA was determined by formaldehyde-agarose gel electrophoresis. Generally, about 40–45 μ g of RNA (1.9 kb) were obtained from 2.5 μ g of DNA template.

The prepared hepsin RNA (1–2 μ g) was then subjected to translation at 30°C in the presence of [³⁵S]methionine by employing the rabbit reticulocyte lysate system (New England Biolab) according to the manufacturer's instructions. An aliquot (5 μ l) of the translation reaction mixture (25 μ l) was mixed with the loading buffer, treated in boiling water for 5 min, and subjected to SDS-polyacrylamide gel (15%) electrophoresis. After electrophoresis, the polyacrylamide gel was treated with Amplify (Amersham) for 15 min according to the manufacturer's instructions to enhance the radioactivity signals, dried, and exposed to x-ray film at -70°C .

RESULTS

Subcellular Localization of Hepsin—Immunoblot analysis of HepG2 as well as BHK cells is shown in Fig. 1. Based on the 5'-nucleotidase activity assayed, the plasma membrane preparation used in this experiment was found to be enriched 18-fold over the crude cell membrane starting material. The membrane preparation was highly pure with little contamination by microsomes and mitochondria, as monitored by glucose 6-phosphatase and succinate-cytochrome c reductase (<0.2% and 0.5% contamination, respectively). Protein bands of 51 and 28 kDa were observed at high concentration levels in the extracts of cell membrane fractions prepared from

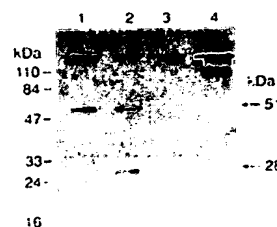


FIG. 1. Immunoblot analysis of HepG2 and BHK cells. Experimental details are described under "Experimental Procedures." Aliquots (7.5 μ g) of proteins of various cell subcomponents and media are loaded for each lane. Lane 1, BHK cell membranes; Lane 2, HepG2 cell membranes; Lane 3, HepG2 cytosol; Lane 4, HepG2 media. The numbers on the left show the positions of size markers.

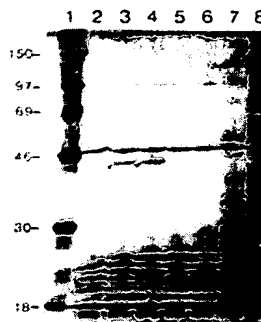


FIG. 2. Cell free translation assay of hepsin cDNA. Lane 1, ¹⁴C-labeled size marker proteins (from the top: myosin, γ -globulins, phosphorylase b, bovine serum albumin, ovalbumin, carbonic anhydrase, lactoglobulin, cytochrome c, respectively); Lane 2, no RNA added; Lanes 3 and 4, 0.42 and 1.7 μ g *in vitro* transcripts (sense strand) were added, respectively; Lanes 5 and 6, 0.42 and 1.7 μ g of *in vitro* transcripts (antisense strand) added, respectively; Lane 7, 1.8 μ g of pSG5 (no hepsin insert) transcribed RNA; Lane 8, 2 μ g of human placenta RNAs. The numbers on the left indicate the positions and sizes of relevant size marker proteins.

HepG2 cells, while BHK cells showed only the major band (51 kDa). These bands were competed out with the addition of P5 (synthetic peptide of the carboxyl-terminal region of hepsin) which was used to raise the antibodies employed in the immunoblot analysis. These bands were also present at reduced levels in nuclear and mitochondrial fractions (data not shown), but neither in the cytosol nor in culture media. The presence of hepsin in nuclei and mitochondria may be due in part to possible cell membrane contamination in these fractions. These results indicate that hepsin is a protein primarily associated with the plasma membrane.

Cell-free Translation Analysis—When *in vitro* transcripts of hepsin cDNA were employed in cell-free translation assays, a specific polypeptide band of 44 kDa was observed in SDS-polyacrylamide electrophoresis (Fig. 2). The estimated size of the band agreed reasonably well with that expected from the cDNA (14). The larger molecular size observed in immunoblot analyses of all extracts may be due to the lack of potential post-translational modifications such as glycosylation. A possible site for the N-linked carbohydrate chain attachment is at amino acid 112 of the hepsin molecule. Hepsin may also contain O-linked carbohydrate chains.

Tissue Distribution of Hepsin Gene Expression—The tissue distribution of hepsin expression was analyzed by RNA blot analysis of total RNA samples prepared from a young adult

baboon tissue including the hypothalamus, small intestine, pancreas, testis, salivary gland, skeletal muscle, lung, adrenal gland, thyroid, pituitary gland, liver, spleen, kidney, brain, and thymus (Fig. 3). The results showed that the mRNA for hepsin was 1.85 kb in size, and was found at the highest level in the liver. It was also present in other tissues, albeit at much lower levels, including the kidney, pancreas, lung, thyroid, pituitary gland, as well as the testis. Extremely low levels of the mRNA were found in the thymus, spleen, small intestine, and in the adrenal gland. These results indicate that hepsin is ubiquitously expressed in various tissues with preferred tissue specificity for liver.

Chromosomal Localization of the Hepsin Gene—To obtain a chromosome assignment for the hepsin gene, a hepsin cDNA probe was hybridized to Southern blots of a panel of somatic cell hybrids. The results showed perfect concordance between human chromosome 19 and hepsin (Table I). A significant discordance was observed between hepsin and all of the other human chromosomes (27–59%).

To determine the chromosomal localization of the hepsin gene using an independent method and to sublocalize this gene, we hybridized a hepsin-specific probe (cDNA) to normal metaphase chromosomes. This resulted in specific labeling only of chromosome 19. Of 100 metaphase cells examined from this hybridization, 39 were labeled on region q1 of one or both chromosome 19 homologues. The distribution of labeled sites on this chromosome is illustrated in Fig. 4. Of 224 total labeled sites observed, 64 (28.6%) were located on chromosome 19. These sites were clustered at bands q11–13.2 and this cluster represented 21.9% (49/224) of all labeled sites (cumulative probability for the Poisson distribution is $\ll 0.0005$). The largest number of grains was observed at 19q13.1. Similar results were obtained in three additional hybridization experiments using this probe. Thus, the hepsin gene is localized to chromosome 19, at bands q11–13.2.

Immunofluorescent Staining of Cultured Cells—Cultured cells including HepG2 and BHK cells were immunostained for hepsin with antibodies (HAbPM) raised against the synthetic peptides (PM, an equimolar mixture of P1, P2, and P3) designed to the catalytic subunit of hepsin (Fig. 5A). The antibodies employed uniformly stained HepG2 cells. BHK cells were also stained, but at reduced intensity. The staining was completely competed out when synthetic peptides used

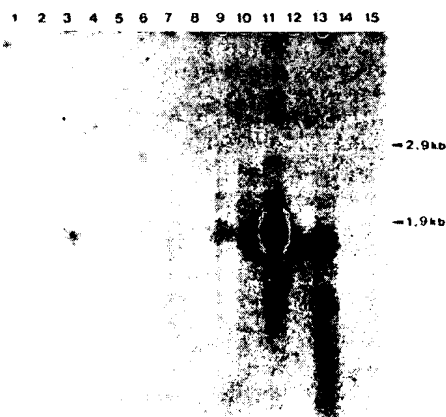


FIG. 3. RNA blot analysis of young adult baboon tissue. Each lane contained 20 μ g of total RNAs isolated from a young adult baboon. Lanes 1–15 contain hypothalamus, small intestine, pancreas, testis, salivary gland, skeletal muscle, lung, adrenal gland, thyroid, pituitary gland, liver, spleen, kidney, brain, and thymus, respectively. The size and positions of RNAs are shown at the right. A hepsin cDNA (1.8 kb) was used as the radiolabeled probe in this experiment.

TABLE I

Synteny test of the hepsin gene and human chromosomes in rodent-human hybrid clones

Somatic cell hybrids were scored for the presence (+) or absence (–) of specific human chromosomes by gene enzyme and cytogenetic analyses and for the presence or absence of hepsin coding sequences by Southern blot hybridization.

Human chromosome	Hepsin gene/human chromosome				Asynteny
	+/+	+/-	-/+	-/-	%
1	7	8	2	20	27
2	4	8	4	10	46
3	2	5	4	12	39
4	6	7	6	9	46
5	5	10	5	17	40
6	12	3	6	12	27
7	3	8	2	12	40
8	5	8	4	6	52
9	5	10	2	15	38
10	5	4	10	11	47
11	10	5	5	16	28
12	9	6	5	10	37
13	5	8	12	9	59
14	10	4	8	11	38
15	7	8	9	12	47
16	9	6	7	11	39
17	10	5	12	8	49
18	6	5	8	6	52
19	15	0	0	22	0
20	9	6	7	6	46
21	5	10	6	15	44
22	3	6	5	11	44
X	10	5	16	6	57

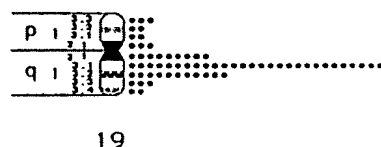


FIG. 4. Distribution of labeled sites on chromosome 19 in 100 normal human metaphase cells from phytohemagglutinin-stimulated peripheral blood lymphocytes that were hybridized with the hepsin probe. Of 64 labeled sites observed on chromosome 19, 49 (76.6%) were clustered at 19q11–13.2; the largest cluster of grains was located at 19q13.1.

for raising antibodies were preincubated with antibodies, indicating that the staining of the cells is specific (Fig. 5B). Antibodies raised against the synthetic peptide P5 (the carboxyl-terminal region) gave similar results (data not shown). Permeabilized cells with Triton X-100 did not show any significant increase or change in staining (data not shown). When antibodies specific for blood coagulation factor IX were used or anti-hepsin antibodies were omitted in control experiments, no significant staining of the cells was observed. These immunostaining patterns show that hepsin primarily has its catalytic subunit (carboxyl-half) at the cell surface. Consequently, its amino-terminal portion is likely to be facing the cytosol. The immunostaining results of cultured cells as well as tissues are consistent with this molecular orientation of hepsin at the cell membrane. These results also agree well with those of the immunoblot analysis which showed hepsin to be primarily located in the cell membrane fraction. The HAbP1 antibody which was raised against the NH₂-terminal region of hepsin did not serve to further confirm the results because of its unfortunate low specificity.

Mild Proteolysis of HepG2 Cells—To further test the topology of hepsin, HepG2 cells were mildly digested with trypsin (100 μ g/ml) or proteinase K (10 μ g/ml) on ice. The results of immunoblot analyses of these protein samples are shown in

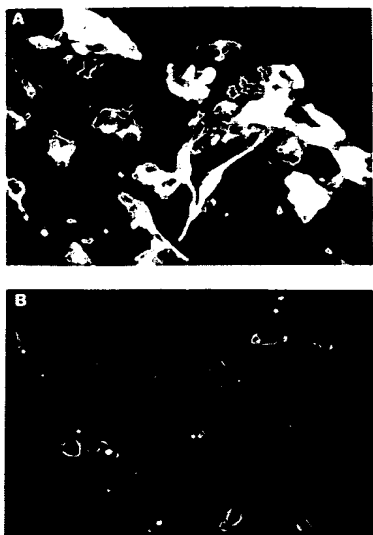


FIG. 5. Fluorescent immunostaining of HepG2 cells. Panel A, staining cells with antibodies raised against the catalytic domain (HABPM). Panel B, staining cells in the presence of antigen peptides. Experimental details are described under "Experimental Procedures."

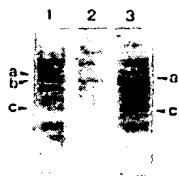


FIG. 6. Immunoblot analysis of plasma-membrane proteins prepared from HepG2 cells with and without mild proteolysis. Lane 1, membrane proteins (20 μ g) of HepG2 cells treated with proteinase K (10 μ g/ml); Lane 2, membrane proteins (20 μ g) of HepG2 cells treated with trypsin (100 μ g/ml); Lane 3, membrane proteins (20 μ g) of HepG2 cells without protease treatment. Bands a and c correspond to the 51- and 28-kDa hepsin bands in Fig. 1. Band b corresponds to partially degraded hepsin. Antibodies prepared against the carboxyl-terminal region (HABP5) were used in this experiment.

Fig. 6. The protein bands (a and c in control lane 3) correspond to 51 and 28-kDa bands of hepsin in Fig. 1. When the cells were treated with trypsin (lane 2), both bands a and c were grossly reduced in intensity compared to the nontreated control (lane 3). When the cells were very mildly treated with proteinase K (10 μ g/ml, lane 1), both bands a and c lowered their intensities and a new band, b, appeared, likely derived from band a. These results suggest that limited proteolysis, which is mild enough to maintain cellular integrity and viability, results in significant degradation of the carboxyl-terminal portion (the catalytic subunit) of hepsin. This further supports the molecular orientation of hepsin with its catalytic subunit at the cell surface exposed to the extracellular space.

DISCUSSION

The results of our studies demonstrate that hepsin, originally identified as a putative membrane-bound protease, is present in the cell membranes. We have also characterized its molecular size, tissue distribution of expression, and the chromosomal localization of its gene.

The size of the mRNA for baboon hepsin is estimated to be about 1.85 kb. The human hepsin mRNA produced in HepG2 has a similar size and agrees well with that predicted from the cDNA. The hepsin gene is located at 19q11-13.2. The

molecular mapping results and Southern blot analysis of human genomic DNA suggest that hepsin has a single copy gene.²

Antibodies raised against synthetic peptides designed to various parts of the hepsin sequence predicted from the cDNA were successfully used to characterize and analyze its expression. Immunoblot analysis of membrane proteins of HepG2 cells showed two polypeptide bands of 51 (major) and 28 kDa (minor) (Fig. 1), whereas BHK cells had only the major band (51 kDa). This major band agrees well with the molecular sizes predicted from the cDNA and the cell-free translation experiment. The smaller minor band of 28 kDa is considered to be a degradation product derived from the putative catalytic subunit portion of the 51-kDa species. In the reduced condition, the apparent size of the 51-kDa band increased slightly indicating that this band represents a single polypeptide chain which has not undergone any degradation during the membrane protein extraction procedures employed. We speculate that proteolytically activated hepsin, which may be composed of two subunits (162 and 255 amino acid residues) linked together with a disulfide bond (14), may be efficiently cleared from the cell membrane, since we have not seen any significant generation of the expected subunits in the gel electrophoretic analysis employing reduced conditions. This may take place by binding to a specific inhibitor(s) or by accelerated degradation due to an unknown mechanism. *In vitro* translation assays of the RNA transcripts of hepsin cDNA showed a distinct specific band of about 44 kDa that agrees reasonably well with the size predicted from the cDNA sequence. This size also agrees well with that observed for cultured cells if we take into account the potential post-translational modifications such as glycosylation which may increase the molecular mass to the apparent 51 kDa. A potential site for the N-linked carbohydrate chain attachment is located at amino acid 112. At the present time, we do not know whether or not this site is glycosylated, or whether any O-linked carbohydrate chains are attached to the mature hepsin molecule.

As shown in RNA blot analysis of baboon tissues (Fig. 3), hepsin appears to be ubiquitously expressed in various tissues, particularly in the liver, at a high level. The expression of hepsin in various tissues suggests that this protease may be involved in an essential biological process(es) in many different cells. In HepG2 cells, hepsin is present in the cell membrane fraction at high levels, but not in the cytosol or in culture media (Fig. 1). Nuclear and mitochondrial fractions also contained a lower amount of hepsin of the same molecular weights (data not shown). The results of fluorescent immunostaining experiments show that hepsin is primarily a cell membrane-associated protease with the molecular orientation of its catalytic subunit (the carboxyl-terminal half) at the cell surface. The patterns of the fluorescent immunostaining of various tissues is consistent with this molecular orientation. The observation that mild protease treatment of intact HepG2 cells greatly decreases the intensity of hepsin bands as tested by immunoblot analysis (Fig. 6) further supports the molecular orientation. When the sequence of 15 amino acid residues which immediately flank the hydrophobic sequence of hepsin were compared, the NH₂-terminal side flanking sequence contained the 4 positive net charges while the COOH-terminal flanking side contained no net charges. This agrees well with the consensus topological sequence for the type II membrane proteins derived from well-defined membrane-spanning proteins (28-30). Furthermore, the immediately flanking residue of the NH₂-terminal side of the hydrophobic sequence is a

² A. Tsuji and K. Kurachi, unpublished data.

positively charged residue, lysine, agreeing well with the consensus sequence for topology of the type II membrane proteins recently proposed by Parks and Lamb (31). These observations support the premise that the mechanism of intracellular transportation of the newly synthesized hepsin is analogous to that of other reported membrane-bound proteins.

Several proteases with a similar cellular localization and orientation have been reported (8, 11, 13). Hepsin, however, is novel and distinct from each of these proteases reported to date.

Proteases have been shown to be present during cell migration (32) and tissue rearrangement (33) involved in morphogenesis, where it has been assumed that they create space for cell migration and process extension through an extracellular matrix and cell-filled milieu. Their role in cell growth can be inferred from their presence, for example, on immature but not mature glial cells (34) or the highly developmentally regulated appearance of tissue plasminogen activator in maturing sperm (35). Although the precise biological role(s) of hepsin is unknown at the present time, we postulate that hepsin also plays an important role(s) in cell growth, probably by creating space for growing cells by degrading a specific extracellular matrix protein(s) or a protein(s) in the tissue. In this regard, it is important to note our recent observation that hepsin is expressed at a greatly elevated level in actively dividing cells in such tissues as the basal layer of the epidermis of developing skin.³ Hepsin may also have a role in other cell functions in normal as well as in pathological conditions. In our preliminary results, antisense oligonucleotides of hepsin show a significant effect on the growth rate as well as on the morphology of HepG2 and BHK cells in culture, supporting the above hypothesis.² Hepsin may also play an important role in the metastasis of tumor tissues like some other membrane proteases (13); however, this has yet to be tested.

Determination of the substrate specificity of hepsin is obviously very important in order to define its precise biological role(s). In our preliminary assay, hepsin highly enriched on the antibody affinity column showed strong activity towards *N*-benzoyl-Leu-Ser-Arg-pNA·HCl, but it did not cleave *N*-benzoyl-Glu-Phe-Ser-Arg-pNA·HCl. To this end, efforts to isolate hepsin in quantity from cultured cells and tissues is in progress. Determination of its concentration in various tumor tissues is also in progress in our laboratory.

Acknowledgments—We thank L. H. J. Lin, Rafael Espinosa III, Matt Rebentisch, and L. Landa for their excellent technical assistance. We also thank Dr. Amiya K. Hajra for his help in assaying the membrane preparations and Dr. Richard E. Tashian for his critical reading of the manuscript. C. Herrerias is also thanked for her excellent help in preparing the manuscript.

REFERENCES

- Neurath, H. (1985) *Fed. Proc.* **44**, 2907–2913
- Bond, J. S., and Butler, P. E. (1987) *Annu. Rev. Biochem.* **56**, 333–364
- Katunuma, N., and Kominami, E. (eds) (1989) *Intracellular Proteolysis. Mechanisms and Regulations. Proceedings of 7th International Committee on Proteases Meetings, Shimoda, Japan*. Japan Scientific Societies Press, Tokyo
- Billings, P. C., Carew, J. A., Keller-McGandy, C. E., Goldberg, A. L., and Kennedy, A. R. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84**, 4801–4805
- Scott, G. K., Seow, H. F., and Tse, C. A. (1989) *Biochim. Biophys. Acta* **1010**, 160–165
- Scott, G. K., and Seow, H. F. (1985) *Exp. Cell Res.* **158**, 41–52
- Fraser, J. D., and Scott, G. K. (1984) *Mol. Immunol.* **21**, 311–320
- Steven, F. S., and Al-Ahmad, R. K. (1983) *Eur. J. Biochem.* **130**, 335–339
- Steven, F. S., Griffin, M. M., and Al-Ahmad, R. K. (1986) *J. Chromatogr.* **376**, 211–219
- Steven, F. S., and Griffin, M. M. (1988) *Biol. Chem. Hoppe-Seyler* **369**, (suppl.) 137–143
- Tanaka, K., Nakamura, T., and Ichihara, A. (1986) *J. Biol. Chem.* **261**, 2610–2615
- Satoh, M., Yukosawa, H., and Ishii, S.-I. (1988) *J. Biochem. (Tokyo)* **103**, 493–498
- Aoyama, A., and Chen, W. T. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8296–8300
- Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* **27**, 1067–1074
- Kurachi, K., Tsuji, A., and O'Shea, K. S. (1989) *Intracellular Proteolysis. Mechanisms and Regulations. Proceedings of 7th International Committee on Proteases Meetings, Shimoda, Japan*. (Katunuma, N., and Kominami, E., eds) pp. 144–149, Japan Scientific Societies Press, Tokyo
- Reichlin, M. (1980) *Methods Enzymol.* **70**(A), 159–165
- Belsham, G. J., Denton, R. M., and Tanner, M. J. A. (1980) *Biochem. J.* **192**, 457–467
- Windell, C. C., and Unkeless, J. C. (1968) *Proc. Natl. Acad. Sci. U. S. A.* **61**, 1050–1057
- Sottocasa, G. L., Kuylenskierna, B., Ernster, L., and Bergstrand, A. (1967) *J. Cell Biol.* **32**, 415–438
- Fleming, P. J., and Hajra, A. K. (1977) *J. Biol. Chem.* **252**, 1663–1672
- Blobel, G., and Potter, V. R. (1966) *Science* **154**, 1662–1665
- Bradford, M. M. (1976) *Anal. Biochem.* **72**, 248–254
- Towbin, H., Staehelin, T., and Gordon, J. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76**, 4350–4354
- Chomczynski, P., and Sacchi, N. (1987) *Anal. Biochem.* **162**, 156–159
- Kusumoto, H., Hirose, S., Salier, J. P., Hagen, F. S., and Kurachi, K. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7307–7311
- Le Beau, M. M., Pettenati, M. J., Lemons, R. S., Diaz, M. O., Westbrook, C. A., Larson, R. A., Sherr, C. J., and Rowley, J. D. (1988) *Cold Spring Harbor Symp. Quant. Biol.* **51**, 899–909
- Le Beau, M. M., Westbrook, C. A., Diaz, M. O., and Rowley, J. D. (1984) *Nature* **312**, 70–71
- Hartmann, E., Rapoport, T. A., and Lodish, H. F. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 5786–5790
- von Heijne, G. (1986) *EMBO J.* **5**, 3021–3027
- Boyd, D., and Beckwith, J. (1990) *Cell* **62**, 1031–1033
- Parks, G. D., and Lamb, R. A. (1991) *Cell* **64**, 777–787
- Valinsky, J. E., and LeDourarin, N. M. (1985) *EMBO J.* **4**, 1403–1406
- Saksela, O., and Rifkin, D. B. (1988) *Annu. Rev. Cell. Biol.* **4**, 93–126
- Kalderon, N., and Williams, C. A. (1986) *Dev. Brain Res.* **25**, 1–9
- Huarte, J., Belin, D., Bosco, D., Sappino, A.-P., and Vassalli, J. D. (1987) *J. Cell Biol.* **104**, 1281–1289

³ S. O'Shea, A. Tsuji, and K. Kurachi, submitted for publication.

Exhibit 38

Identification and Cloning of the Membrane-associated Serine Protease, Hepsin, from Mouse Preimplantation Embryos*

(Received for publication, August 5, 1997, and in revised form, October 2, 1997)

Thien-Khai H. Vu†§, Rose W. Liu‡, Carol J. Haaksma¶, James J. Tomasek||, and Eric W. Howard‡

From the Departments of ‡Pathology and ¶Anatomical Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73190 and the ||Center for Assisted Reproductive Technology, Columbia-Presbyterian Hospital, Oklahoma City, Oklahoma 73104

Previous studies have suggested the existence of a membrane-associated serine protease expressed by mammalian preimplantation embryos. In this study, we have identified hepsin, a type II transmembrane serine protease, in early mouse blastocysts. Mouse hepsin was highly homologous to the previously identified human and rat cDNAs. Two isoforms, differing in their cytoplasmic domains, were detected. The tissue distribution of mouse hepsin was similar to that seen in humans, with prominent expression in liver and kidney. In mouse embryos, hepsin expression was observed in the two-cell stage, reached a maximal level at the early blastocyst stage, and decreased subsequent to blastocyst hatching. Expression of a soluble form of hepsin revealed its ability to autoactivate in a concentration-dependent manner. Catalytically inactive soluble hepsin was unable to autoactivate. These results suggest that hepsin may be the first serine protease expressed during mammalian development, making its ability to autoactivate critical to its function.

Embryonic development is marked by a series of cellular divisions and morphogenetic changes (1). These processes are mediated by the complex expression and interplay of different sets of genes, some of which are derived from maternally expressed genes stored as mRNAs in the oocytes. It is generally accepted that zygotic gene expression begins at the embryonic two-cell stage (2). These newly expressed zygotic genes complement the maternally expressed genes to mediate early preimplantation development. Numerous studies have suggested the involvement of a variety of proteases during development. Members of the astacin family of metalloproteases are involved in hatching in both invertebrates and vertebrates (3–6), pattern and tentacle cell formation in the hydra by HMP1 (7), neuroblast migration in *Caenorhabditis elegans* by hch-1 (3), dorsal/ventral patterning in *Drosophila* by Tolloid (8), and biomineralization and bone/cartilage formation in mammals by BMP-1 (9, 10). Interestingly, both Tolloid and BMP-1 can physically interact with transforming growth factor- β (8, 9), and this association is essential for normal development, perhaps to activate latent transforming growth factor- β complexes. In ad-

dition, BMP-1 has been shown to be the procollagen C-endopeptidase (EC 3.4.24.19) required for the processing of type I, II, and III procollagen to fibrillar collagens to yield the major fibrous components of vertebrate extracellular matrix (11, 12).

Proteases have also been shown to play essential roles in cell differentiation. Recently, new members of the adamalysin/reprolysin metalloprotease have been described and were shown to have a direct role in a number of developmental processes. Fertilin- α and - β , the first members of this family, have been shown to have essential roles in sperm-egg fusion during fertilization (13–15). The recent discovery of meltrin- α , a fertilin-related member of the adamalysin/reprolysin metalloproteases important for myoblast fusion during skeletal muscle development, suggests that there may be a common mechanism in gamete and myoblast fusion (16). Astacin-like proteases of the Tolloid/BMP-1 family play important roles in cell differentiation and morphogenesis in animal embryos ranging from the hydra and sea urchins to mammals (17).

Serine proteases have also been implicated in development, which is exemplified by genetic studies of the products of the *Drosophila* gene *stubble-stubloid*, which is essential for epithelial morphogenesis of imaginal discs of *Drosophila* (18). Mutations in this gene affects imaginal disc formation and affect the organization of microfilament bundles, leading up to defects in bristle, leg, and wing morphogenesis. Also in *Drosophila*, the maternally transcribed product of the *easter* gene, a trypsin-like serine protease, is essential for the establishment of a normal dorsal-ventral pattern in the embryos (19). Of note, perturbing quantitatively the level of Easter protease activity in *Drosophila* as a result of dominant mutations can disrupt the dorsal-ventral axis, leading to ventralizing and lateralizing phenotypes (20). The *Drosophila* trypsin-like enzymes easter and snake are part of a cascade of zymogen activation leading up to the conversion of the ligand-precursor, spatzle to its active form (21–23). Active spatzle then activates its receptor Toll to affect specification of dorsal and lateral cell fates (24, 25).

While evidence exists that one or more serine proteases exist in mammalian preimplantation embryos, the identity of these enzymes has remained elusive. One of the earliest events in embryogenesis thought to require a protease is blastocyst hatching. This involves the proteolysis of the zona pellucida, an event critical for subsequent uterine implantation of the embryo. Studies have suggested that a single, membrane-associated serine protease is expressed by hatching blastocysts (26). In this study, we identify hepsin, a serine protease containing a transmembrane domain, as a serine protease expressed by mouse embryos at the two-cell stage through the early blastocyst stage. In addition, we demonstrate that a soluble form of hepsin lacking the transmembrane domain undergoes autoactivation, suggesting a mechanism by which hepsin becomes proteolytically activated in the absence of other proteases.

* This work was supported by the Oklahoma Center for the Advancement of Science and Technology Grant HR4-098. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact. The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AF030065.

§ To whom correspondence should be addressed: Dept. of Pathology, Biomedical Sciences Bldg., Rm. 434, University of Oklahoma, Health Sciences Center, 940 Stanton L. Young Blvd., Oklahoma City, OK 73190. Tel.: 405-271-1483; Fax: 405-271-8774.

1	GCTGGCTGCTGCTGCCACCTTTCCTGCTCCGGGCTGCTCCGCTGCTGGGAGACAGACACCATGCCCTGCCAGGCCGGAGACTAAC	CGGAGGAGGGGGCTCCGACGGCCACGCT
31	ATG AAG AAG GAG GGT GGC CGG ACT GCA GCA TGC TGC TCC AGA CCC AAG GTG GCA GCT CTC ATT GTG	
118	CCCAAACTGACCACTCTCCGGCGAACCCAGGGTTCGGCCCGACCCAAAGGTCAACCTGGGAATCATTAACAAGAGTCCCTGAC	
205	M A K E G G R T A A C C S R P K V A A L I V	22
	ATG AAG AAG GAG GGT GGC CGG ACT GCA GCA TGC TGC TCC AGA CCC AAG GTG GCA GCT CTC ATT GTG	
271	G T L L F L T G I G A A S M A I V T I L L O	44
	GGT ACC CTG CTG TTC CTG ACA GGC ATT GGG GCC GCG TCC TGG GCC ATT GTG ACC ATC CTA CTG CAG	
337	S D Q E P L Y Q L S P G D S R L A V L G	66
	AGT GAC CAG GAG CCA CTG TAC CAA GTG CAG CTC AGT CCA GGG GAC TCA CGA CTT GCA GTG TTG GAC	
403	K A T E G T W R L L C S S R S N A R V A G G L G	88
	AAG ACG GAG GGT ACG TGG AGG CTA CTG TGC TCC TCA CGC TCC AAT GCC AGG GTG GCA GG CTC GGC	
469	C E E M G F L R A L A H S E L D V R T A G A	110
	TGT GAG GAG ATG GGC TTT CTC AGG GCT CTG GCG CAC TCG GAG CTG GAT GTG CGC ACT GCG GGC GCC	
535	N G T S G F F C V D G G L P L A Q R L L G	132
	AAC GGC ACA TCG GGC TTC TTT TGC GTG GAC GAG GGC GGA CTG CCT CTG GCT CAG AGG TTG CTG GAT	
601	V I S V C D C P F R G R F L T A T C C Q D C G R	154
	GTC ATC TCT GTA TGT GAC TGT CCA AGA GGC CGA TTC CTG ACT GCC ACC TGC CAA GAC TGT GGC CGC	
667	R K L P V D R I V G G Q D S S L G R W P W Q	176
	AGG AAG CTG CCG GTG GAC CGC ATT GTG GGG GGC CAG GAC AGC AGT CTG GGA AGG TGG CCG TGG CAG	
733	V S L R Y D G T H L C G G S L L S G D W V L	198
	GTC AGC CTG CGT TAT GAT GGG ACC CAC CTC TGT GGG GGG TCC CTG CTG TCT GGG GAC TGG GTG CTG	
799	T A A H C F P A R N R V L S R W R V F A G A	220
	ACT GCT GCA CAT TGC TTT CCA GAG CGG AAC CGG GTC CTG TCT CGG TGG CGA GTA TTT GCT GGT GCT	
865	V A R T S P H A V Q L G V Q A V I Y H G G Y	242
	GTA GCC CGG ACC TCA CCC CAT GCT GTG CAA CTG GGG GTT CAG GCT GTG ATC TAT CAT GGG GGC TAC	
931	L P F R D P T I D E N S N D I A L V H L S S	264
	CTT CCC TTT CGA GAC CCT ACT ATT GAC GAA AAC AGC AAT GAC ATT GCC TTG GTC CAC CTC TCT AGC	
997	S L P L T E Y I Q P A A G Q A L V G D	286
	TCC CTG CCT CTC ACA GAA TAC ATC CAG CCA GTG TGT CTC CCT GCT GCG GGA CAG GCC CTG GTG GAT	
1063	G K V C T V T G W G N A C Q F Y G Q Q A M V L	308
	GCC AAG GTC TGT ACT GTG ACC GGC TGG GCT AAC ACA CAG TTC TAT GGC CAA CAG GCT ATG GTG CTC	
1129	C A E A R V P I I S N E V C N S P D F Y G N Q	330
	CAA GAG GCC CGG GTT CCC ATC ATA AGC AAC GAA GTT TGC AAC AGC CCC GAC TTC TAC GGG AAT CAG	
1195	I K P K M P C A G Y P E G G I D A C Q G G A C	352
	ATC AAG CCC AAG ATG TTC TGT GCT GGC TAT CCT GAG GGT GGC ATT GAT GCG TGC CAG GGC GAC AGT	
1261	G G P P V C E D A G A G C A T C T G G A C A A G G T T G G C G A C T A G G T T G G C T C C A C A C T	374
	GGA GGC CCC TTT GTG TGT GAA GAC AGC ATC TCT GGG ACA TCA AGG TGG CGG CTA TGT GGC ATT GTA	
1327	S W G T G C A L A R K P G V Y T K V T D F R	396
	AGC TGG GGT ACC GGC TGT GCT TTG GCC CGG AAG CCA GGA GTG TAC ACC AAA GTC ACT GAC TTC CGG	
1393	E W I F K A I K T H S E A S G M V T Q P Stop	416
	GAG TGG ATC TTC AAG GCC ATA AAG ACT CAC TCC GAA GCC AGT GGC ATG GTG ACT CAG CCC TGA TCC	
1459	CGCCTCATCTCGCTGCTCCGCTGCTGCACTAGCATCCAGAGTCAGAGTTGGTCTGGTGGCTCCAGCCCACTGGTAGGCTCCACACT	
1546	GGGCTTCACATGGAATGGTTCTCTGCTCAGATCCAGTCCAGGGTCCAAAGGATGCTGGATCCAAAGGACTTCTTCCACAGTGGCCG	
1633	GCCCACTCAATCCAGGGCCATGGGCTCACCTCCACCCCATGTAAATATTACTCTGCTCTGGGGGGCGCTCTAGGGAGCCCC	
1720	TTGTGCAGATGCTCTTTAAATAATAAAGGTGGTTTGGATTATGGGAGAAAAA	

FIG. 1. Nucleotide and predicted amino acid sequences of mouse hepsin. The internal signal peptide sequence serving as a transmembrane (TM) domain is underlined. The zymogen activation cleavage site (arrow), catalytic triad residues (asterisks), and Asp³⁴⁶ (circle) are depicted.

EXPERIMENTAL PROCEDURES

Collection and Culture of Mouse Preimplantation Embryos—Experiments utilizing preimplantation embryos were performed with cultured two-cell stage embryos, which were obtained from B6C3F1 prepubescent female mice (Charles Rivers Lab) weighing 10–13 g. Mice were injected intraperitoneally with 5 IU of pregnant mare's serum gonadotropin (Sigma) followed 48 h later with 5 IU of human chorionic gonadotropin (Sigma). Subsequently, a single female was paired with a single male overnight, and females were checked for vaginal plugs the following day (day 1). On day 2, mice were dissected to obtain the oviducts, which were bathed in sperm washing medium (Irvine Scientific) and dissected to release the two-cell embryos. About 40–50 two-cell embryos were pooled and cultured under oil at 37 °C in a humidified atmosphere of 5% CO₂ in air in 50-μl droplets of human tubular fluid (Irvine Scientific) plus 0.5% human serum albumin (Irvine Scientific). Cultures were maintained for 4–5 days or until expanded blastocysts began to hatch.

RNA Isolation and First-strand cDNA Synthesis—Total RNA was isolated from 100–200 hatching blastocysts (embryonic day 4.5), according to the method of Chomczynski and Sacchi (27). The total amount of RNA obtained was then used in the first-strand cDNA synthesis reaction using SuperScript reverse transcriptase (Life Technologies, Inc.) and oligo(dT) as primers. The reaction was incubated at 42 °C for 1 h. Subsequently, RNase H (Life Technologies, Inc.) was added and the reaction was incubated at 37 °C for 20 min to remove the RNA template.

PCR Amplification, Cloning, and Sequencing of Mouse Hepsin—To

identify the serine protease involved in mouse blastocyst hatching, degenerate oligonucleotides, 5'-TGCTCTAGATGG(A/G)TINTI(A/T)(G/C)IGCIGCICA-3' and 5'-CCGGAATTCA(A/G)IGGI(G/C)(ACT)ICCI(G/C)(A/T)(A/G)TCICC-3' (Molecular Biology Resource Facility, OUHSC), based on two conserved regions of known serine proteases, were used to amplify a 500-bp DNA fragment, encoding part of the protease catalytic domain, from hatching blastocyst RNA. Aliquots of first-strand cDNA were incubated in the presence of 0.1 μM of each 5'- and 3'-primers, 100 μM dNTP, 1 × PCR buffer, and 2.5 units/100 μl of AmpliTaq DNA polymerase (Perkin-Elmer). The reactions were cycled 40 times through the following steps: 30 s at 94 °C, 30 s at 55 °C, and 1 min at 72 °C in a Perkin-Elmer DNA thermocycler model 2800. DNA fragments of the correct size (~500 bp) were purified from agarose gels using GeneClean II (BIO 101 Inc., Vista, CA). The purified fragments were ligated into pBS-SK+ (Stratagene) using T4 DNA ligase (New England Biolabs). Double-stranded DNA was sequenced using T3 and T7 primers and the Sequenase Version 1 kit (U. S. Biochemical Corp./Amersham Life Science). Sequences of cloned PCR fragments were compared with DNA sequences compiled in data bases.

A full-length cDNA of mouse hepsin was subsequently cloned by screening a mouse liver cDNA library (Stratagene), using the manufacturer's instruction. ³²P-Labeled DNA probes were generated using the Prime-It II random primer labeling kit (Stratagene) and the 500-bp cloned PCR fragment described above as a template. A 1.8-kb cDNA obtained was sequenced as described above using both pBluescript and internal primers.

Construction and Expression of Soluble Hepsin and Catalytically Inactive Hepsin—The method of site-directed mutagenesis as described previously (28, 29) was used to introduce a *StuI* restriction site at the end of the coding sequence of the transmembrane domain of hepsin using the oligonucleotide, 5'-GTGACCATCCTAAGGCCTAGTGAC-CAGGAGCC-3', which replaced nucleotides 331–336 with a *StuI* site.

¹ The abbreviations used are: PCR, polymerase chain reaction; bp, base pair(s); kb, kilobase pair(s); RT, reverse transcriptase; PAGE, polymerase chain reaction; fVII, factor VII; fVIIa, activated factor VII; pBS, pBluescript.

Mouse	MAKEGGRTAACCSRFK	VAALIVGTLFLFTGIGAASWAIVTILL	QSDQEPLYQVQLSPG	58
Rat	-----AP-----	---TV---F---G-----IL-	R-----Q---L-PG	58
Human	MAQ-----VP-----	---TA---L---A-----AV-	R-----P---V-SA	59
Mouse	DSRLAVLDKTEGTWRLC	SSRSNARVAGLGCEMGFLRALAHSELDVRTAGANGTSGFFC		118
Rat	-S--L-L-----	-----G-----A-----		118
Human	-A--M-F-----	-----S-----T-----		119
Mouse	VDEGGLPLAQRLLDVIS	VCDCPRGRFLTATCQDCGRKLPVDRIVGGQDSSLGRWPQVS		178
Rat	---G--LA---D-----	---T-T-----Q-S-----		178
Human	---R--HT---E-----	---A-I-----R-T-----		179
Mouse	LYRDGTHLCGGSLLSGD	WVLTAAHCFPERNRVLSRWRVFAGAVARTSPHAVQLGVQAVIY		238
Rat	-----T-----	-----RT--AV--I--		238
Human	-----A-----	-----QA--GL--V--		239
Mouse	HGGYLPFRDPTIDENS	NDIALVHLSSSLPLTEYIQPVCLPAAGQALVDGKVCTVTGCGWT		298
Rat	-----TID-----	-----S-----V-----		298
Human	-----NSE-----	-----P-----I-----		299
Mouse	QFYQQAMVLQEARVPII	SNEVCNSPDFYGNQIKPKMFCAGYPEGGI	D ACQGDSCGPF	356
Rat	-F---V-----E---SP---		-----H---	356
Human	-Y---G-----D---GA---		-----P---	357
Mouse	VCEDSISGTSRWRLCGI	VSWGTCALARKPGVYTKVTDFFREWIFKAIKTHSEASGMVTQP*		417
Rat	---R--G-S-----	---R-----I-----Q-----T-----P*		417
Human	---S--R-P-----	---Q-----S-----Q-----S-----L*		418

FIG. 2. Sequence alignment of mouse, rat, and human hepsin. Deduced amino acid sequences of mouse, rat, and human hepsin are shown. Amino acid identity is indicated by a dash. The conserved TM domain and Asp³⁴⁶ are boxed.

This *StuI* site and the *XbaI* site at the 3' end of the cDNA in pBS-SK+ were used to excise a 1.1-kb DNA fragment and cloned into the same sites in the RSV-PL4 expression vector (30). This construct included a transferrin signal peptide, followed by an amino-terminal epitope tag recognized by HPC4, a calcium-dependent monoclonal antibody (31). The soluble hepsin expressed using this vector had a new amino-terminal of Glu-Asp-Gln-Val-Asp-Pro-Arg-Leu-Ile-Asp-Gly-Lys-Ile-Glu-Gly-Ser-Pro, followed by the wild-type hepsin sequence from Ser⁴⁶. The non-functional S348A soluble hepsin mutant, which replaced the active site serine with an alanine, was constructed similarly with the additional use of the oligonucleotide, 5'-TGCCAGGGCGACGCTGGGGCCCTTTGTG-3'. The resulting constructs were transfected into human 293 epithelial cells using LipofectAMINE (Life Technologies, Inc.) as suggested by the manufacturer. High expressing clones were selected using 400 µg/ml G418 (Life Technologies, Inc.). The accuracy of the constructs were confirmed by DNA sequencing. The recombinant epitope-tagged protein was purified from conditioned medium by affinity chromatography using HPC4-linked Affi-Gel 10 and was eluted with EDTA.

Assay of Soluble Hepsin Activity—Soluble hepsin amidolytic activity was assayed using the chromogenic substrate Spectrozyme PCA (H-D-[Cbo]-Lys-Pro-Arg-pNA; American Diagnostica) at a final concentration of 0.2 mM. The absorbance at 405 nm was monitored over 10 min using a V_{max} microplate reader (Molecular Devices) to determine the rate of chromogenic substrate hydrolysis ($\Delta A_{405}/\text{min}$). Inhibitory dose-response curves were generated by preincubating the enzyme with specific inhibitors at different concentrations for 30 min at ambient temperature prior to the addition of the substrate.

Semiquantitative RT-PCR and Southern Blot Analysis—RT-PCR-linked Southern blot analysis to augment sensitivity of detection was utilized to investigate the temporal expression of hepsin in mouse preimplantation embryos. cDNAs from various stages of development were prepared from 40 to 50 embryos as described above. Oocytes were prepared from unmated females and treated with hyaluronidase (Sigma) to remove cumulus cells before proceeding to the total RNA isolation and cDNA synthesis as above. PCR was performed essentially as above with the mouse hepsin primers, 5'-ATCCAGCCAGTGTGTCTCCTG-3' and 5'-TCAGGGCTGAGTCACCATGCCAC-3', but with only 15 cycles. Similar PCR reactions using, β -actin primers (a gift from Jeff Gimble, Department of Surgery, University of Oklahoma Health Sciences Center), were used as positive controls. Southern blot analysis of the PCR products was performed as described previously (30) using ³²P-labeled random-primed DNA probes generated from the same amplified DNA regions as templates.

Northern Blot Analysis—Total RNA was isolated from cells according to published methods (27). RNA was transferred to MSI-NT nylon membranes by capillary action, then cross-linked to membranes with UV light. Membranes were incubated for 1 h at 60 °C with prehybridization buffer (500 mM NaPO₄, pH 7.4, 7% SDS, 1 mM EDTA). Membranes were then hybridized overnight in prehybridization buffer plus labeled cDNA probe at 60 °C. Probes were ³²P-labeled by random prim-

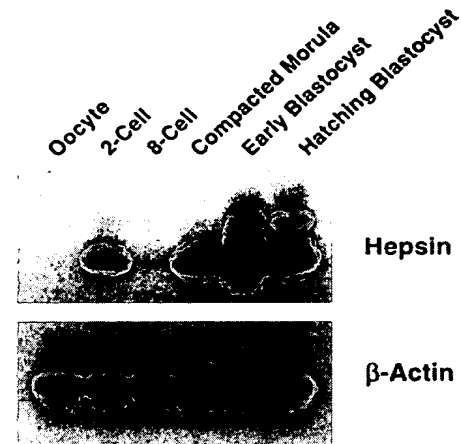


FIG. 3. Temporal expression of hepsin in mouse preimplantation embryos. Total RNA from mouse embryos was isolated, then analyzed for hepsin mRNA expression by Southern blot-linked-RT-PCR analysis ($n = 3$). β -Actin was used as a control.

ing using a Prime-it II kit (Stratagene), then separated from unincorporated label using ProbeQuant G-50 Micro columns (Pharmacia Biotech). Following three low stringency washes (15 min in 40 mM NaPO₄, pH 7.2, 5% SDS, 1 mM EDTA, 0.5% bovine serum albumin at room temperature), and two high stringency washes (15 min with 40 mM NaPO₄, pH 7.2, 1% SDS, 1 mM EDTA at 60 °C), and one 30-min high-stringency wash, membranes were exposed to x-ray film adjacent to an enhancing screen.

RESULTS

Strategy for the Identification and Cloning of an Embryonic Serine Protease—A prior study using a radioiodinated active site chloromethyl ketone probe and SDS-PAGE detected a single serine protease of $M_r = 74,000$ in mouse blastocyst lysates (26). Using RT-PCR and degenerate oligonucleotides based on conserved regions in the catalytic domain of serine proteases, we amplified and subcloned a 0.5-kb cDNA fragment encoding the putative mouse hatching enzyme from hatching blastocysts mRNAs. Ten separate clones were sequenced and found to be identical. Data base searches showed that the deduced amino acid sequence was similar to that of human hepsin, a trypsin-like serine protease previously cloned from a liver library (33). A full-length mouse hepsin cDNA (Fig. 1) was obtained after

screening a mouse liver library using the amplified DNA fragment as a probe. Hepsin is a type II transmembrane protein with an extracellular carboxyl-terminal catalytic domain (33, 34). Based on the predicted amino acid sequence homology with other related serine proteases, hepsin is likely to be synthesized as a single chain zymogen that requires cleavage of the Arg¹⁶¹-Ile¹⁶² bond to generate the mature, disulfide-linked two-chain form. In addition to the catalytic triad residues and Asp³⁴⁶, which is important for trypsin-like specificity, the transmembrane and short cytoplasmic domains of hepsin are all conserved among mouse, rat, and human hepsin (Fig. 2). The significance of the transmembrane domain remains to be determined.

Temporal Expression of Hepsin in Preimplantation Embryos—To determine if the temporal expression of hepsin was consistent with that of a hatching enzyme, we performed semi-quantitative RT-PCR-linked Southern blotting to indirectly determine the time and level of hepsin message in oocytes and in several stages of preimplantation development. Hepsin transcription was biphasic, beginning at the 2-cell stage, absent at the 8-cell stage, and peaking at the early blastocyst stage prior to hatching (Fig. 3). There was no detectable expression in oocytes, and, subsequent to embryo hatching, the level of expression clearly diminished (Fig. 3).

Tissue Expression and Multiple Hepsin mRNAs—Human hepsin was previously shown to be expressed primarily in liver and kidney, and mouse hepsin was similarly distributed (Fig. 4). Unlike human hepsin, mouse hepsin had two alternative forms detected by Northern blotting, migrating at 1.8 and 1.9 kb. To characterize the differences in the two hepsin mRNAs, we performed RT-PCR analysis using total RNA samples isolated from mouse liver and kidney. Several oligonucleotide primers spanning the hepsin cDNA sequence were utilized, as shown in Fig. 5. PCR analysis revealed that an insert in the

5'-end of the coding sequence distinguished the 1.9-kb message from the 1.8-kb message. DNA sequencing revealed an additional 60-bp sequence coding for 20 amino acids within the cytoplasmic domain of 1.9-kb hepsin cDNA (Fig. 6). This sequence has not been demonstrated in human hepsin.

Expression and Autoactivation of Soluble Hepsin—Because hepsin is a type II transmembrane serine protease, we wanted to address the possibility that a soluble form of the enzyme could be expressed and used to elucidate hepsin's enzymatic properties. We developed an expression construct by site-directed mutagenesis that encoded for a zymogen form of hepsin lacking its transmembrane and cytoplasmic domains (soluble hepsin), and stably expressed it in human 293 epithelial cells. Soluble hepsin was expected to be expressed as a single-chain zymogen which could be activated proteolytically to a disulfide-linked two-chain form, consisting of a 12-kDa light chain and 31-kDa heavy chain. The intact precursor as well as proteolytically activated species would be expected to migrate with a M_r = 43,000 on SDS-PAGE gels. Surprisingly, upon elution, soluble hepsin was spontaneously activated from a single-chain zymogen to the active disulfide-linked two-chain form (Fig. 7, WT lanes, and data not shown); this activation was not detected in the conditioned medium not subjected to purification (Fig. 7,

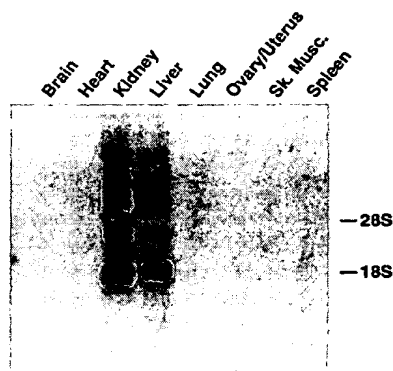


FIG. 4. Tissue distribution of mouse hepsin expression. Total RNA (20 μ g/lane) from several adult rat tissues was analyzed for hepsin expression by Northern blots hybridized with a cDNA consisting of the entire hepsin coding region. Two hybridizing species highly detected in liver and kidney correspond to mRNAs of approximately 1.8 and 1.9 kb in size.

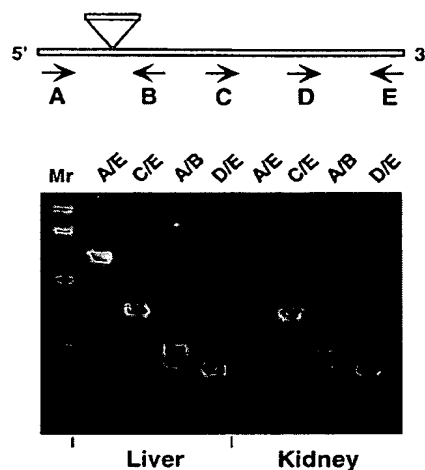


FIG. 5. Localization of the region of nucleotide insertion in the 1.9-kb hepsin message. Total RNA from both mouse kidney and liver were subjected to RT-PCR analysis using different primers sets (each primer is denoted by a letter from A-E) to localize the region of nucleotide differences between the 1.8- and 1.9-kb hepsin mRNAs. The positions of the primers (arrows) are indicated along the 5'- to 3'-nucleotide sequence as represented by a horizontal bar above the gel image. The position of the nucleotide insertion is also marked. PCR products were separated by 1% agarose electrophoresis and stained with ethidium bromide. Primer set A/B detected two different bands due to the 60-bp insertion in the coding region for the cytoplasmic domain of hepsin. Primers were as follows: A, 5'-TGGGAATCATTAACAA-GAGTCCCTGAC-3'; B, 5'-AGTCAGGAATCGGCCTCTAGG-3'; C, 5'-AGGAAGCTGCCGGTGGACCGCATTGTG-3'; D, 5'-ATCCAGCCAGTGTGTCTCCCTG-3'; E, 5'-TCAGGGCTGAGTCACCATGCCAC-3'.

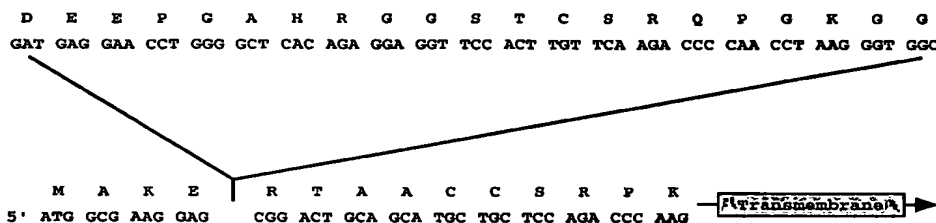


FIG. 6. Alternative cytoplasmic domains in the two hepsin mRNAs. Amino acid and cDNA sequence of the hepsin cytoplasmic domain, with the inserted sequence within the 1.9-kb form shown above the 1.8-kb form of hepsin.

CM lanes). Additionally, it further processed itself from a 43- to 29-kDa form (Fig. 7, *non-reduced WT lane*). Upon reduction, only a 31-kDa band, which represented the heavy or catalytic chain, was seen, suggesting that only the light chain was proteolytically modified to generate the 29-kDa form seen under nonreducing conditions. The autoactivation of soluble hepsin upon elution was not seen with a catalytically inactive S352A soluble hepsin mutant, in which the active site serine was replaced by alanine (Fig. 7, *S352A lanes*). Of note, the initial eluate, when immediately prepared and separated by reducing SDS-PAGE, showed only a small amount of conversion to the two-chain form (data not shown). Similarly, the presence of the inhibitor benzamidine in the eluate prevented the conversion and only a small converted fraction was seen on reducing SDS-PAGE (data not shown).

DISCUSSION

We have identified hepsin, a membrane-bound serine protease previously shown to activate fVII (35), in preimplantation mouse embryos as early as the two-cell stage. Based on evidence that a single serine protease is present in preimplantation embryos (26), it is possible that hepsin represents the first such protease expressed during development. Prior *in vitro* experimentation implicated hepsin in the maintenance of cellular morphology and hepatoma cell growth (36), and in blood coagulation by human factor VII activation (35). Increased hepsin expression has also been associated with ovarian cancer (37). No developmental functions of hepsin have been described. Whether hepsin plays a critical role in early development is not clear, but it is possible that it plays a role in blastocyst hatching.

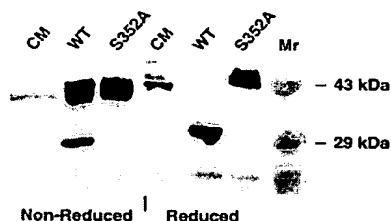


FIG. 7. Soluble hepsin is capable of autoactivation. Wild-type and S352A soluble hepsin was isolated from medium conditioned by transfected 293 epithelial cells, and proteins were separated by both nonreducing and reducing SDS-PAGE and blotted to nitrocellulose membrane. The primary HFP-2 and anti-goat alkaline phosphatase-conjugated antibodies were used to visualize hepsin in conditioned medium (CM), as well as purified soluble hepsin (WT) and its inactive mutant (S352A). Molecular mass markers are shown in kDa.

The hepsin amino acid sequence suggests it is a type II transmembrane serine protease zymogen with an extracellular carboxyl-terminal catalytic domain. The internal signal sequence, serving as a transmembrane domain, is surprisingly conserved. The presence of this transmembrane domain is consistent with Perona and Wassarman's (26) data suggesting that the putative mouse hatching enzyme, which would be expressed in early preimplantation embryos, is membrane-bound. The trypsin-specificity conferring Asp³⁴⁶ that lines the S1 subsite and composes part of the specificity pocket is present and conserved, indicating that hepsin is likely to have trypsin-like specificity. Indeed, our activity assay of the recombinant soluble hepsin using a number of chromogenic substrates have confirmed this observation. The reason for the presence of two forms of hepsin, differing in the cytoplasmic domain, is not clear. The inserted sequence in the 1.9-kb form of hepsin has no homology to any domains found in signal transducing proteins. It is unlikely that changes to the cytoplasmic domain alter hepsin's proteolytic properties, particularly since the soluble form of the enzyme is apparently fully functional. Whether the 1.8- and 1.9-kb hepsin mRNAs are the result of two different genes or, more likely, the result of alternative splicing of a single gene transcript remains to be defined.

Since hepsin is likely to be expressed as a zymogen based on the predicted amino acid sequence, and appears to be the only serine protease present during blastocyst hatching, the question arises, what is the mechanism of its activation? Our hypothesis is that density-dependent autoactivation occurs, as suggested by data from our soluble hepsin expression study. We noted that during purification, upon elution with EDTA, soluble hepsin was spontaneously converted to the active, disulfide-linked two-chain form probably via cleavage of the Arg¹⁶¹-Ile¹⁶² bond. The conversion was clearly concentration dependent (activation was only seen in the eluate and not in the diluted conditioned medium) and required hepsin's inherent enzymatic activity since it was not observed with a catalytically inactive S352A mutant soluble hepsin. These data indicate that hepsin was capable of concentration-dependent autoactivation. Since hepsin is membrane-bound via a transmembrane domain, its density and lateral diffusion on the trophoblast surface may play an important role in achieving the concentration needed for autoactivation (Fig. 8). This mode of autoactivation resembles fVII cell surface autoactivation, which utilizes distinct tissue factor molecules to localize both the fVII and fVIIa to the cell surface, forming two separate membrane-bound binary complexes. The complex with the active fVIIa then activates the adjacent tissue factor-anchoring

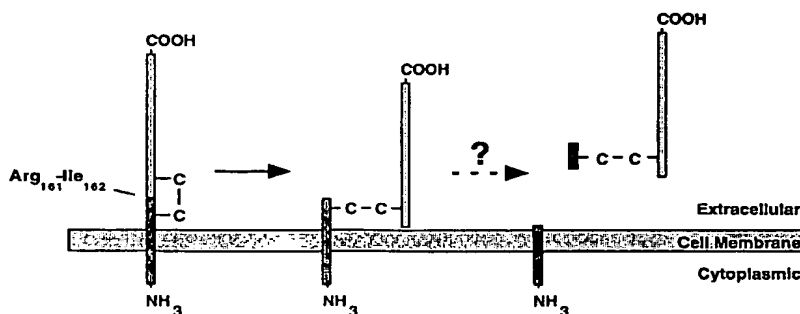


FIG. 8. Model of hepsin activation. Based on structural similarities to other serine proteases, hepsin is expressed as a single-chain zymogen and can be activated proteolytically by a single cleavage at the Arg¹⁶¹-Ile¹⁶² bond to generate the two-chain, membrane-bound form. Its deduced primary amino acid sequence suggests that hepsin is expressed as a type II transmembrane zymogen with an extracellular carboxyl catalytic domain. The heavy or catalytic chain is linked to the light chain via a disulfide bond (C-C). The light chain is anchored to the cell membrane by a hydrophobic, internal signal sequence. Based on the soluble hepsin expression studies, the mode of activation on the cell surface is likely to be autoactivation. Our evidence further suggests that a soluble form, resulting from additional cleavages of the membrane-bound light chain, is possible.

fVII, obeying obligatory two-dimensional enzyme kinetics (38). Hepsin autoactivation is likely to follow similar kinetics, but further studies are necessary to elucidate its mechanism of cell surface autoactivation. Interestingly the recent purification of intact hepsin from rat liver microsomes also resulted in its activation (39), but it was not clear if this was the result of autoactivation or of the action of another protease. Our data with the inactive hepsin mutant suggest that membrane-bound hepsin is capable of autoactivation.

The autoactivation of soluble hepsin additionally generated a second form of the enzyme. A band of 29 kDa, which was absent in the S352A mutant, along with the intact 43 kDa, were both present when the eluate was analyzed on nonreducing SDS-PAGE and Western blot experiments. This 29-kDa form was likely to be the result of proteolytic modification of the light chain of the active two-chain form since only the intact catalytic heavy chain was seen under reducing conditions. The presence of this 29-kDa form suggests that membrane-bound hepsin can be cleaved off the trophoblast surfaces of embryos (Fig. 8). Interestingly, Sawada *et al.* (40) have demonstrated the presence of a soluble trypsin-like activity in blastocyst culture medium and that this activity represented that of a hatching enzyme. Whether this secreted trypsin-like activity and the 29-kDa form of hepsin are one and the same, and what roles it may play during embryogenesis, remain to be determined.

Acknowledgments—We thank Dr. Charles Esmon for financial support in making the antibody HFP-2, and the generous gifts of the mAb HPC4 and HPC4-linked Affi-Gel 10. We also thank Dr. Alireza Rezaie for the use of RSV-PL4, the V_{max} microplate reader (Molecular Devices), and the chromogenic substrate Spectrozyme PCa (American Diagnostica).

REFERENCES

- McLaren, A. (1982) *Reproduction in Mammals*, Cambridge University Press, Cambridge, MA
- Nothias, J. Y., Majumder, S., Keneko, K. J., and DePamphilis, M. L. (1995) *J. Biol. Chem.* **270**, 22077–22080
- Hishida, R., Ishihara, T., Kondo, K., and Katsura, I. (1996) *EMBO J.* **15**, 4111–4122
- Yasumasu, S., Yamada, K., Akasaka, K., Mitsunaga, K., Iuchi, I., Shimada, H., and Yamagami, K. (1992) *Dev. Biol.* **153**, 250–258
- Elaroussi, M. A., and DeLuca, H. F. (1994) *Biochim. Biophys. Acta* **1217**, 1–8
- Katagiri, C., Maeda, R., Yamashika, C., Mita, K., Sargent, T. D., and Yasumasu, S. (1997) *Int. J. Dev. Biol.* **41**, 19–25
- Yan, L., Pollock, G. H., Nagase, H., and Sarraz, M. P., Jr. (1995) *Development* **121**, 1591–1602
- Shimell, M. J., Ferguson, E. L., Childs, S. R., and O'Connor, M. B. (1991) *Cell* **67**, 469–481
- Wozney, J. M., Rosen, V., Celeste, A. J., Mitscock, L. M., Whitters, M. J., Kriz, R. W., Hewick, R. M., and Wang, E. A. (1988) *Science* **242**, 1528–1534
- Fukagawa, M., Suzuki, N., Hogan, B. L., and Jones, C. M. (1994) *Dev. Biol.* **163**, 175–183
- Kessler, E., Takahara, K., Biniaminov, L., Brusel, M., and Greenspan, D. S. (1996) *Science* **271**, 360–362
- Li, S. W., Sieron, A. L., Fertala, A., Hojima, Y., Arnold, W. V., and Prockop, D. J. (1996) *Proc. Natl. Acad. Sci. U. S. A.* **93**, 5127–5130
- Wolfsberg, T. G., Bazan, J. F., Blobel, C. P., Myles, D. G., Primakoff, P., and White, J. M. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90**, 10783–10787
- Myles, D. G., Kimmel, L. H., Blobel, C. P., White, J. M., and Primakoff, P. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91**, 4195–4198
- Evans, J. P., Schultz, R. M., and Kopf, G. S. (1995) *J. Cell. Sci.* **108**, 3267–3278
- Huovila, A. P. J., Almeida, E. A., and White, J. M. (1996) *Curr. Opin. Cell. Biol.* **8**, 692–699
- Bond, J. S., and Beynon, R. J. (1995) *Protein Sci.* **4**, 1247–1261
- Appel, L. F., Prout, M., Abu-Shumays, R., Hammonds, A., Garbe, J. C., Fristrom, D., and Fristrom, J. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90**, 4937–4941
- Chasan, R., and Anderson, K. V. (1989) *Cell* **56**, 391–400
- Jin, Y., and Anderson, K. V. (1990) *Cell* **60**, 873–881
- Stein, D., and Nusslein-Volhard, C. (1992) *Cell* **68**, 429–440
- Schneider, D. S., Jin, Y., Morisato, D., and Anderson, K. V. (1994) *Development* **120**, 1243–1250
- Smith, C. L., and DeLotto, R. (1994) *Nature* **368**, 548–551
- Morisato, D., and Anderson, K. V. (1994) *Cell* **76**, 677–688
- Morisato, D., and Anderson, K. V. (1995) *Annu. Rev. Genet.* **29**, 371–399
- Perona, R. M., and Wassarman, P. M. (1986) *Dev. Biol.* **114**, 42–52
- Chomczynski, P., and Sacchi, N. (1987) *Anal. Biochem.* **162**, 156–159
- Vu, T.-K. H., Hung, D. T., Wheaton, V. I., and Coughlin, S. R. (1991) *Cell* **64**, 1057–1068
- Vu, T.-K. H., Wheaton, V. I., Hung, D. T., Charo, I., and Coughlin, S. R. (1991) *Nature* **353**, 674–677
- Rezaie, A. R., and Esmon, C. T. (1992) *J. Biol. Chem.* **267**, 26104–26109
- Stearns, D. J., Kurosawa, S., Sims, P. J., Esmon, N. L., and Esmon, C. T. (1988) *J. Biol. Chem.* **263**, 826–832
- Deleted in proof
- Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* **27**, 1067–1074
- Tsuji, A., Torres-Rosado, A., Arai, T., Le Beau, M. M., Lemons, R. S., Chou, S.-H., and Kurachi, K. (1991) *J. Biol. Chem.* **266**, 16948–16953
- Kazama, Y., Hamamoto, T., Foster, D. C., and Kisiel, W. (1995) *J. Biol. Chem.* **270**, 66–72
- Torres-Rosado, A., O'Shea, K. S., Tsuji, A., Chou, S.-H., and Kurachi, K. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7181–7185
- Tanimoto, H., Yan, Y., Clarke, J., Korourian, S., Shigemasa, K., Parmley, T. H., Parham, G. P., and O'Brien, T. J. (1997) *Cancer Res.* **57**, 2884–2887
- Neuenschwander, P. F., Fiore, M. M., and Morrissey, J. H. (1993) *J. Biol. Chem.* **268**, 21489–21492
- Zhukov, A., Hellman, U., and Ingelman-Sundberg, M. (1997) *Biochim. Biophys. Acta* **1337**, 85–95
- Sawada, H., Yamazaki, K., and Hoshi, M. (1990) *J. Exp. Zool.* **253**, 83–87

Exhibit 39

A Novel Transmembrane Serine Protease (TMPRSS3) Overexpressed in Pancreatic Cancer^{1,2}

Christine Wallrapp, Susanne Hähnel, Friederike Müller-Pillasch, Beata Burghardt, Takeshi Iwamura, Manuel Ruthenbürger, Markus M. Lerch, Guido Adler, and Thomas M. Gress³

Department of Internal Medicine I, University of Ulm, 89081 Ulm, Germany [C. W., S. H., F. M.-P., G. A., T. M. G.]; Hungarian Academy of Sciences, University of Budapest, 1450 Budapest, Hungary [B. B.]; Department of Medicine B, University of Münster, 48129 Münster, Germany [M. M. L., M. R.]; Miyazaki Medical College, Kiyotake, Miyazaki 889-1692, Japan [T. I.]

Abstract

We report the characterization of a novel serine protease of the chymotrypsin family, recently isolated by cDNA-representational difference analysis, as a gene overexpressed in pancreatic cancer. The 2.3-kb mRNA of the gene, named *TMPRSS3*, is strongly expressed in a subset of pancreatic cancer and various other cancer tissues, and its expression correlates with the metastatic potential of the clonal SUIT-2 pancreatic cancer cell lines. The deduced polypeptide sequence consists of 437 amino acids and exhibits all of the structural features characteristic of serine proteases with trypsin-like activity. *TMPRSS3* is membrane bound with a NH₂-terminal signal-anchor sequence and a glycosylated extracellular region containing the serine protease domain. Thus, *TMPRSS3* is a novel membrane-bound serine protease overexpressed in cancer, which may be of importance for processes involved in metastasis formation and tumor invasion.

Introduction

Proteases have been increasingly recognized as important factors in the pathophysiology of tumorous diseases. The proteolytic degradation of the extracellular matrix, which is an indispensable step in tumor invasion and metastasis, is mediated by members of the four major classes of endopeptidases, including serine, cysteine, aspartyl, and metalloproteases (1). In this highly complicated process, a cascade of events requiring a variety of proteases seems to be involved. Numerous reports have demonstrated an increased production of extracellular matrix degrading enzymes, including type IV collagenase (MMP-2), cathepsin B, cathepsin D, and serine proteases such as plasminogen activator in tumor cells (1). The proteolytic enzymes of the serine protease family exist as single-chain or double-chain zymogens activated by specific and limited proteolytic cleavage. They contain the three active-site amino acids histidine, aspartate, and serine, which participate in peptide bond hydrolysis. The geometric orientation of this catalytic triad is similar in different serine proteases, despite the fact that folding of the proteases may be different (2).

In the present study, we report the cloning and characterization of a novel serine protease identified in a recent cDNA-RDA⁴ approach (3). This study was designed to isolate gene fragments highly over-

expressed in pancreatic cancer compared with normal pancreas and chronic pancreatitis tissue. From the 16 gene fragments isolated in this study, we selected the 313-bp gene fragment RDA12 (GenBank accession no. U54603) for further characterization. Database comparison revealed a moderate homology to a number of serine proteases, indicating that RDA12 may be a fragment of a novel protease with cancer-specific expression.

Materials and Methods

Materials. Human tissue from patients with ductal adenocarcinoma of the pancreas ($n = 13$), carcinoma tissues of different origin, human pancreatic tissue from organ donors ($n = 6$), and chronic pancreatitis tissue ($n = 6$) was provided by the Hungarian Academy of Sciences (Budapest, Hungary) and the Department of Surgery of the University of Ulm. All tissue samples were obtained after approval by the local Ethics Committee.

The human pancreatic cancer cell lines were obtained from the following suppliers: PATU-8988S and PATU-8988T (German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany); PANC-1 and MIA-PaCa-2 (European Collection of Animal Cell Cultures, Salisbury, United Kingdom); HPAF (Metzgar, Durham, NC); Capan-1, Capan-2, and AsPC-1 (Cell Lines Service, Heidelberg, Germany); Patu II (Elsässer, Marburg, Germany); PC2 (Bülow, Mainz, Germany); SUIT-2 (S2-007, S2-013, S2-020, and S2-028; Iwamura, Miyazaki, Japan; Ref. 4); and SKPC2 and IMIM-PC2 (P. Real, IMIM, Barcelona, Spain).

Cloning of a New Serine Protease cDNA. In a recent screen for differentially expressed genes in pancreatic carcinoma, the 313-bp gene fragment RDA12 (accession no. U54603) was isolated by cDNA-RDA (3); this fragment encodes the putative motif of a new serine protease. The RDA12 fragment was used to screen ~20,000 clones of an oligo(dT)-primed cDNA library from a pancreatic cancer cell line by hybridization. Both strands of the longest cDNA clone, RDA12/2, were sequenced by primer walking. For stable transfection in mammalian cells, the cDNA clone RDA12/2 was cloned in sense and antisense orientation into the *Bam*HI site of the mammalian expression vector pHB-Ap1-neo (5). A COOH-terminal-tagged *TMPRSS3* expression vector was constructed by insertion of a 1427-bp fragment (nucleotides 96–1522) containing the open reading frame of *TMPRSS3* into the *Bst*XI site of the mammalian expression vector pcDNA6/V5/His B (Invitrogen, San Diego, CA).

Northern Blot Analyses. The expression of *TMPRSS3* was studied by hybridizations using Northern blots containing 30 µg each of total RNA from normal pancreas tissue, chronic pancreatitis tissue, different carcinoma tissues, and cell lines. The Northern blots containing RNA of different human tissues were purchased from Clontech (Heidelberg, Germany).

Cell Culture and Transfection. For functional analysis of *TMPRSS3*, the S2-020 pancreatic cancer cell line, which expresses no endogenous *TMPRSS3* mRNA, was transfected with the *TMPRSS3*-pHB-Ap1-neo construct in sense and antisense orientation using DMRIE-C (Life Technologies, Inc., Eggenstein, Germany). Several clones were picked that showed various degrees of stable *TMPRSS3* sense/antisense mRNA expression. Two of each sense and antisense clones were used for functional assays.

HEK-293 cells were plated at 1.5×10^6 cells/10-cm dish and grown overnight in DMEM supplemented with 10% FCS. Cells were transiently transfected with the *TMPRSS3*-pcDNA6/V5/His plasmid DNA by use of the calcium phosphate protocol.

Received 12/17/99; accepted 3/30/00.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¹This work was supported by grants from the Bundesministerium für Bildung und Forschung (01 GB9401), the European Community (BMH4-CT98-3085), and the Deutsche Forschungsgemeinschaft (SFB518, project B1; to T. M. G.).

²The nucleotide sequence in this report has been submitted to the GenBank Data Library with accession no. AF179224.

³To whom requests for reprints should be addressed, at Department of Internal Medicine I, University of Ulm, 89081 Ulm, Germany. Phone: 49-731-5024385; Fax: 49-731-5024302; E-mail: thomas.gress@medizin.uni-ulm.de.

⁴The abbreviations used are: RDA, representational difference analysis; PNGase F, peptide-N-glycosidase F.

1
 35 AGGATGCTGGGCGTGAGGGACCAAGGCCTGCCCTGCACTCGGGCCTCTCCAGCCAGTGTGACCAAGGACTTCTGACCTGCTGGCCAGC
 125 CAGGACCTGTGTGGGGAGGCCCTCTGCTGCCTTGGGGTGACAATCTCAGCTCCAGGCTACAGGGAGACCGGGAGGATCAGAGCCAGC
 215 ATGTTACAGGATCTGACAGTGATCAACCTCTGAACAGCCTCGATGTCAAACCCCTGCGCAAAACCCGTATCCCCATGGAGACCTTCAGA
 1 M L Q D P D S D Q P L N S L D V K P L R K P R I P M E T F R
 305 AAGTGGGGATCCCCATCATATAGCACTACTGAGCCTGGCGAGTATCATATTGTGGTTGTCTCATCAAGGTGATTCTGGATAAATAC
 31 K V G T E I I I A L L S L A S I I I V V V L I K V I L D K Y
 395 TACTTCTCTGCGGGCAGCCTCTCCACTTCATCCCGAGGAAGCAGCTGTGTGACGGAGAGCTGGACTGTCCCTTGGGGGAGGACGAGGAG
 61 Y F L C G Q P L H F I P R K Q L C D G E L D C P L G E D E E
 485 CACTGTGTCAAGAGCTTCCCGAAGGGCTGCGAGTGGCAGTCCGCTCTCCAAGGACCGATCCACACTGCAGGTGTGGACTCGGCCACA
 91 H C V K S F P E G P A V A V R L S K D R S T L Q V L D S A T
 575 GGGAACTGGTTCTCTGCCTGTTTCGACAACCTTCAGAAAGCTCTCGCTGAGACAGCCTGTAGGCAGATGGGTACAGCAGCAAAACCACT
 121 G N W F S A C F D N F T E A L A E T A C R Q M G Y S S K P T
 665 TTCAGAGCTGTGGAGATTGGCCAGACCAGGATCTGGATGTTGTTGAAATCACAGAAAACAGCCAGGAGCTTCGCATGCGGAACTCAAGT
 151 F R A V E I G P D Q D L D V V E I T E N S Q E L R M R N S S
 755 GGGCCCTGTCTCTCAGGCTCCCTGGTCTCCCTGCACTGTCTTGCTGTGGGAAGAGCCTGAAGACCCCGTGTGGTGGGTGGGGAGGAG
 181 G P C L S G S L V S L H C L A C G K S L K T P R V V G G E E
 845 GCCTCTGTGGATTCTTGGCCTTGGCAGGTGAGTACGACAAACAGCAGCTGTGTGGAGGGAGCATCTGGACCCCACTGGGTG
 211 A S V D S W P W Q V S I Q Y D K Q H V C G G S I L D P H W V
 935 CTCACGGCAGCCCACTGCTTCAGGAAACATACCGATGTGTTCAACTGGAAGGTGCGGGCAGGCTCAGACAACTGGGCAGCTTCCCATCC
 241 L T A A H C F R K H T D V F N W K V R A G S D K L G S F P S
 1025 CTGGCTGTGGCCAAAGATCATCATTTGAATTAACCCCATGTATCCCAAGACAATGACATCGCCCTCATGAAGCTGCAGTTCCCACTC
 271 L A V A K I I I I E F N P M Y P K D N D I A L M K L Q F P L
 1115 ACTTCTCAGGCACAGTCAGGCCCATCTGTCTGCCCTTCTTTGATGAGGAGCTCACTCCAGCCACCCCACTCTGGATCATTTGGATGGGGC
 301 T F S G T V R P I C L P F F D E E L T P A T P L W I I G W G
 1205 TTTACGAAGCAGAATGGAGGAAGATGTCTGACATACTGCTGCAGGCGTCAGTCCAGGTGATTGACAGCACACGGTGCAATGCAGACGAT
 331 F T K Q N G G K M S D I L L Q A S V Q V I D S T R C N A D D
 1295 GCGTACCAGGGGGAAGTCACCGAGAAGATGTGTGTCAGGCGATCCCGGAAGGGGGTGTGGACACCTGCCAGGGTGACAGTGGTGGGGCC
 361 A Y Q G E V T E K M M C A G I P - E G G V D T C Q G D S G G P
 1385 CTGATGTACCAATCTGACAGTGGCATGTGGTGGGCATCGTTAGCTGGGGCTATGGCTGCGGGGGCCCGAGCACCCAGGAGTATACACC
 391 L M Y Q S D Q W H V V G I V S W G Y G C G G P S T P G V Y T
 1475 AAGGTCTCAGCCTATCTCAACTGGATCTACAATGTCTGGAAGGCTGAGCTGTAATGCTGCTGCCCTTTGCAGTGCTGGGAGCCGCTTCC
 421 K V S A Y L N W I Y N V W K A E L *
 1565 TTCTGCCCTGCCACCTGGGGATCCCCAAAGTCAGACACAGAGCAAGAGTCCCTTGGGTACACCCCTCTGCCACAGCCTCAGCATT
 1655 TCTTGGAGCAGCAAGGGCCTCAATTCTTATAAGGAACCTCGCAGCCAGAGGGCGCCAGAGGAAGTCAGCAGCCCTAGCTCGGCCACA
 1745 CTGGTGTCTCCAGCATCCAGGGAGAGACACGCCCACTGAACAAGGTCTCAGGGGTATTGCTAAGCCAAGAAGGAAGTCTCCCACT
 1835 ACTGAATGGAAGCAGGCTGTCTGTAAAGCCAGATCACTGTGGCTGGAGAGGAGAAGGAAGGGTCTGCCAGCCCTGTCCGTTT
 1925 CACCCATCCCCAAGCCTACTAGAGCAAGAAACAGTGTGAATATAAATGCACTGCCCTACTGTTGGTATGACTACCGTTACCTACTGTT
 2015 GTCATTGTTATTACAGTATGGCCACTATTATTAAGAGAGCTGTGAACATTTCTGGCAAAAAA

Fig. 1. Nucleotide sequence of the cDNA coding for human *TMPRSS3* and its predicted amino acid sequence. The **bold** nucleotide sequence 1189–1501 represents the initially isolated RDA12 gene fragment, the underlined nucleotides 2045–2050 mark the potential polyadenylation signal. The amino acid sequence highlighted by a *gray box* represents the potential transmembrane domain. ▲ indicates the active-site residues histidine (H), aspartate (D), and serine (S). Double underlines indicate potential N-linked glycosylation sites.

Preparation of Cell Extracts and Subcellular Fractionation. Forty-eight h after transient transfection with V5-tagged *TMPRSS3* into HEK-293 cells, protein extracts were prepared by resuspending pelleted cells in 1% Triton X-100, 1% sodium deoxycholate, 0.1% SDS, 150 mM NaCl, 50 mM Tris-HCl (pH 7.2) supplemented with 5 μ g/ml Aprotinin, 5 mM Pefabloc, and 10 μ g/ml Pepstatin. For immunopurification of the epitope-tagged protein, cell lysates were incubated with V5 antibody conjugated to protein G-agarose beads at 4°C for 4 h on a shaker. The agarose beads were pelleted by centrifugation and washed twice with 150 mM NaCl, 5 mM EDTA, 50 mM Tris + 0.1% NP40. The washed pellets were resuspended in 150 mM NaCl, 5 mM EDTA, 50 mM Tris + 0.1% NP40 for PNGase F treatment.

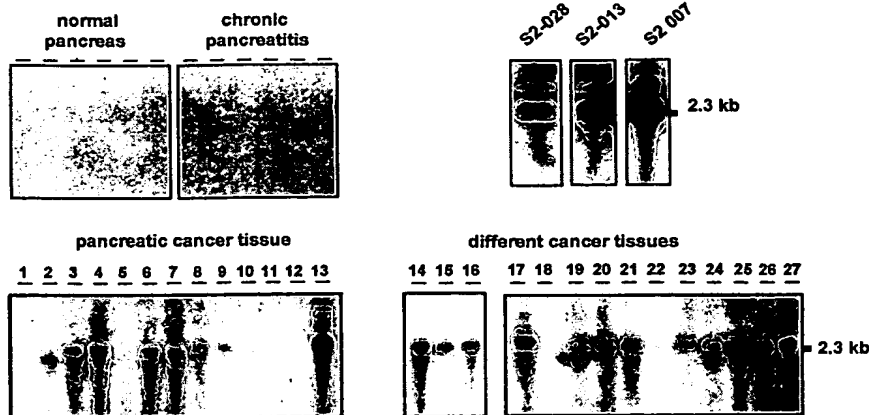
Subcellular fractions were prepared from transiently transfected HEK-293 cells as reported previously (6). The plasma membrane-enriched fraction, which was prepared using sucrose density gradient centrifugation, the cytosolic fraction, and concentrated culture medium were studied by Western blot analysis.

Glycosylation. For PNGase F treatment, immunopurified protein was incubated overnight with 2 units of PNGase F supplemented with 10 mM EDTA at 37°C. Inhibition of N- and mucin-like O-glycosylation was performed by cultivating *TMPRSS3*-expressing HEK-293 cells for 24 h in DMEM, 10% FCS containing either 2.5 μ g/ml tunicamycin (7) or 2 mM phenyl-N-Acetyl- α -D-galactosaminide (8). Thereafter, cells were harvested for protein extraction.

Functional Assays. Nude mouse experiments were done by injecting 2×10^6 S2-020 cells stably transfected with *TMPRSS3* sense/antisense constructs, both s.c. and in the tail vein of female *nu/nu* mice. Five weeks after the tail vein injections, the lung, spleen, and liver were used for standard histological analysis to identify the presence or absence of metastatic lesions. Subcutaneous tumors were measured and used for histological analysis.

In vitro matrigel invasion assays were done by seeding 10^5 transfected cells in medium + 1% FCS in the upper chamber of Matrigel-coated 8- μ m transwell plates. The lower chamber was filled with medium + 10% FCS. The

Fig. 2. Northern blot analyses of the *TMPRSS3* transcript in different tissues and cell lines. The Northern blots contain 30 μ g of total RNA per lane from normal human pancreas ($n = 6$), chronic pancreatitis tissue ($n = 6$), pancreatic carcinoma tissue ($n = 13$; Lanes 1–13), and cancer tissues of different origin (Lanes 14–16, 19–21, and 23, colorectal carcinoma; Lanes 17 and 25–27, gastric cancer; Lane 22, soft tissue sarcoma; Lane 18, breast cancer; Lane 24, carcinoma of the papilla vateri) and the SUI-2 subclones S2-028, S2-013, and S2-007. RNAs from normal pancreas, chronic pancreatitis, and pancreatic cancer tissue samples were run on the same Northern blot gels. The autoradiographs for cancer and control tissues are shown separately for improved presentation of the data.



number of invading cells adhering to the lower side of the porous membrane was counted after fixation with 4% paraformaldehyde and staining with methylene blue.

The proteolytic activity in *TMPRSS3* sense/antisense-transfected S2-020 cells and transiently transfected HEK-293 cells was determined fluorometrically in native lysates and lysates treated with enterokinase for activation, using oligopeptide substrates for elastase-like (Ala-Ala-Ala-Ala) and trypsin-like (Ile-Pro-Arg) serine proteases as described previously (9).

Chromosomal Mapping of the *TMPRSS3* Gene Locus. The chromosomal localization of *TMPRSS3* was determined by screening the GeneBridge4 radiation hybrid panel (Research Genetics, Huntsville, AL), using the *TMPRSS3*-specific primers 5'-CATGTGGTGGGCATCGTTA-3' and 5'-CCAGTTGAGATAGGCTGAG-3'.

Results and Discussion

The 313-bp fragment encoding the putative motif of a new serine protease isolated in a recent cDNA-RDA screen for genes differentially expressed in pancreatic cancer (3) was used to screen a pancreatic cancer cDNA library. Among 16 isolated homologous clones, a clone designated RDA12/2 contained the full-length sequence. The sequence of clone RDA12/2 comprised 2071 bp, including a 214-bp 5' untranslated region, an open reading frame of 1311 nucleotides, and a 546-bp 3' untranslated region (Fig. 1). Translation of the open reading frame suggests that the cDNA codes for a putative polypeptide of 437 amino acids with an estimated molecular mass of 48.202 kDa. The NH₂-terminal region of the hypothetical protein contains a putative signal-anchor sequence characteristic for group II integral membrane proteins. The highly hydrophobic region of 22 amino acids may serve as a transmembrane domain that is involved in anchoring the protease to the cell membrane. According to the charge difference rule (10), it can be assumed that the COOH terminus of the protein with its protease module is located on the extracellular surface.

Although the nucleotide sequence is unique, database comparisons of the amino acid sequence revealed a homology to a number of serine proteases. Thirty-five percent identity and ~50% similarity was found to members of the serine protease family known as the human transmembrane proteases, *TMPRSS1/hepsin* (11) or *TMPRSS2* (12). Thus, our new protease is the third member of a family of transmembrane-bound serine proteases. Consequently, this new gene was named *TMPRSS3* for transmembrane protease, serine 3. Sequence homology was high in the domains containing the three principal active-site amino acids H²⁴⁵, D²⁹⁰, and S³⁸⁷, required for peptide bond hydrolysis. The arrangement of the catalytic residues in the linear sequence defines the membership of *TMPRSS3* to the S1 family of the chymotrypsin clan SA of serine-type peptidases (2). The prototype of this family is chymotrypsin, and the three-dimensional structures of some of its members have already been resolved (12).

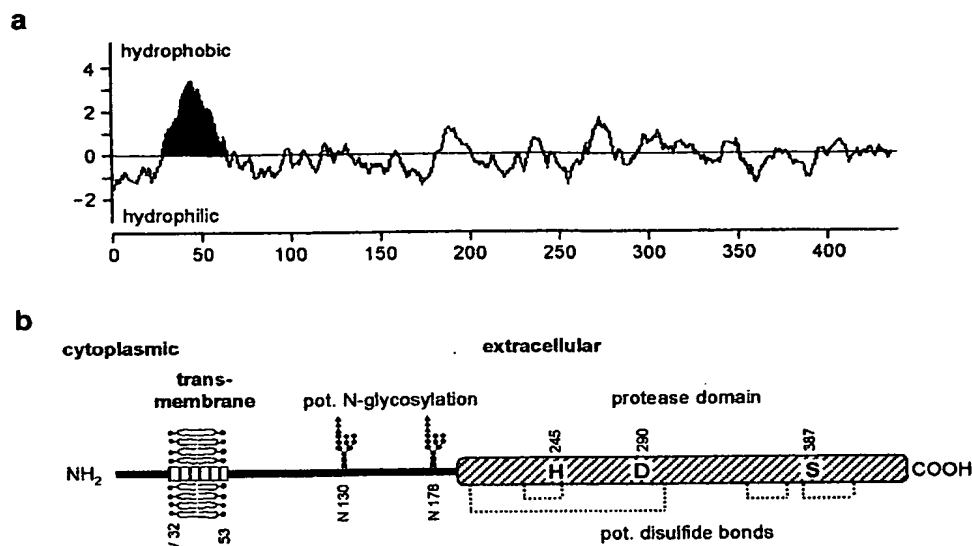
TMPRSS3 is predicted to cleave in a trypsin-like manner after lysine or arginine residues because it contains D³⁸¹ at the base of the specificity pocket that binds the substrate (13). In addition, the novel protein shares considerable structural similarities of the *TMPRSS* family, including the putative NH₂-terminal membrane anchor and the conserved cysteine residues, which by homology most likely form the disulfide bonds C¹⁹⁶-C³¹⁰, C²³⁰-C²⁴⁶, C³⁵⁶-C³⁷², and C³⁸³-C⁴¹⁰. Serine proteases are most commonly synthesized as inactive proenzymes, which are activated by extracellular, proteolytic removal of a propeptide. At the NH₂-terminal part of the protease domain, *TMPRSS3* contains the peptide sequence RVVGG, which is typical for the proteolytic activator site of many protease zymogens. The potential cleavage between R²⁰⁴ and V²⁰⁵ would result in a new terminal α -amino group, which forms a salt bridge with D³⁸⁶ and thereby leads to the assembly of the functional catalytic sites. Therefore, the activated form would consist of a non-protease and a protease subunit linked by a disulfide bond that most likely involves C¹⁹⁶-C³¹⁰. Whether this activation is mediated under physiological conditions by autocatalytic cleavage or other proteases is not known. The *TMPRSS3* gene locus was localized to chromosome 11 at q23.3 between the markers D11S4362 and D11S4387 by use of a radiation hybrid panel.

As anticipated, an overexpression of the 2.3-kb transcript was found in 9 of 13 primary pancreatic carcinoma tissues (Fig. 2) and in 10 of 16 pancreatic carcinoma cell lines (not shown) by Northern blot analysis. Because *TMPRSS3* was not expressed in normal pancreas ($n = 6$) and in chronic pancreatitis ($n = 6$) tissue samples, overexpression appears to be cancer-specific and not due to inflammatory alterations in the stroma. No clear correlation was found between the stage of pancreatic tumors and the expression of the protease (Table 1). Northern blot analyses with RNA from a small number of other tumor tissues revealed that *TMPRSS3* overexpression is not restricted

Table 1. TNM classification of pancreatic cancer patients

Tissue sample	TNM classification
1	T ₃ N ₁ M ₀
2	T ₃ N ₁ M ₀
3	T ₂ N ₁ M ₀
4	T ₂ N ₀ M _x
5	T ₃ N ₁ M _x
6	T ₂ N ₁ M ₀
7	T ₂ N ₁ M ₀
8	T ₂ N ₀ M ₀
9	T ₃ N ₁ M ₀
10	T ₃ N ₁ M ₀
11	T ₃ N ₁ M ₀
12	T ₄ N ₀ M ₁
13	T ₂ N ₁ M ₁

Fig. 3. *a*, hydropathicity plot of the predicted TMPRSS3 protein. The method of Kyte and Doolittle (20) was used, using a window of 17 residues (<http://bioinformatics.weizmann.ac.il/hydroph/>). The peak spanning amino acids 32–53 represents the putative transmembrane domain. *b*, schematic representation of the different domains of TMPRSS3, a type II membrane-associated serine protease. Numbers correspond to the amino acids, deduced from the cDNA sequence shown in Fig. 1. The disulfide bonds were deduced based on the structure of TMPRSS1 and TMPRSS2, the most homologous proteins. *pot.*, potential.



to pancreatic cancer, but can also be found in gastric ($n = 4$), colorectal ($n = 7$), and ampullary ($n = 1$) cancer. No expression was found in one tissue sample each of soft tissue sarcoma and breast cancer (Fig. 2). *TMPRSS3* transcripts were not detectable in normal heart, brain, placenta, lung, liver, skeletal muscle, uterus, and adipose tissue. A weak signal was found in tissues of the normal gastrointestinal tract (esophagus, stomach, small intestine, colon) and in some tissues of the urogenital tract (kidney and bladder). Nevertheless, expression was much weaker than in the corresponding tumors (data not shown). Furthermore, we analyzed the expression of *TMPRSS3* in the SUI-2 clonal cell lines S2-007, S2-013, and S2-028 (4). These subclones of the human pancreatic cancer cell line SUI-2 differ in their spontaneous metastatic potential after s.c. injection in nude mice. In this setting S2-007 regularly shows a high rate of metastases, whereas the other two cell lines show a lower rate (S2-013) or no metastases at all (S2-028). As shown in Fig. 2, the strength of *TMPRSS3* expression correlated well to the metastatic potential of the SUI-2 subclones, which may serve as an indication that this serine protease is associated with the promotion of metastasis.

The sequence of *TMPRSS3* suggests that this novel serine protease contains a signal anchor characteristic for group II integral membrane proteins with a hydrophobic transmembrane domain (Fig. 3*a*). According to the charge difference rule (10), the transmembrane domain (amino acids 32–53) anchors the protease to the cell membrane. Because of this anchorage, the NH_2 -terminal domain (amino acids 1–31) would appear to be located intracellularly, and the COOH-terminal region (amino acids 54–437), which contains the catalytic domain, would be located extracellularly (Fig. 3*b*). The alleged subcellular localization of the protease was confirmed using a V5-tagged *TMPRSS3* construct, which was transiently transfected into HEK-293 cells. Membrane fractionation and Western blotting with the corresponding anti-V5 antibody revealed a signal only in the plasma membrane-enriched fraction, whereas no tagged *TMPRSS3* protein was detectable in the cytosol and in the culture medium (Fig. 4).

This experiment also uncovered post-translational modifications of *TMPRSS3*. Although the calculated theoretical molecular mass of the epitope-tagged fusion protein is 52 kDa, its size in a SDS-polyacrylamide gel is ~68 kDa, suggesting the presence of potential carbohydrate moieties. The primary sequence of *TMPRSS3* displays two consensus motifs for *N*-linked glycosylation (N-X-T/S) at N¹³⁰ and N¹⁷⁸. To confirm this *N*-glycosylation, epitope-tagged *TMPRSS3* was

expressed in HEK-293 cells, immunoprecipitated, and treated with PNGase F. This resulted in an increase in mobility on denaturing SDS-PAGE, demonstrating *N*-glycosylation of *TMPRSS3* (Fig. 4). Cultivation of transfected HEK-293 cells in the presence of tunicamycin, an inhibitor of *N*-glycosylation, revealed the same mobility shift of *TMPRSS3* to a molecular mass of 60 kDa. Phenyl-*N*-acetyl- α -D-galactosaminide, which inhibits mucin-like *O*-glycosylation, had no effect on the molecular mass (data not shown). The generation of recombinant proteases frequently has been shown to be difficult or impossible (14). Despite extensive and repeated efforts, we were unable to successfully generate recombinant protein in *Escherichia coli* and insect cells, possibly because *TMPRSS3*, as many other proteases, had a cytotoxic effect on transfected cells. Repeated efforts to generate peptide antisera failed as well (data not shown), and a *TMPRSS3* antibody was therefore not available for further studies.

Whereas the established physiological role of the chymotrypsin family of secreted serine proteases is primarily in protein catabolism, the function of serine proteases of the *TMPRSS* family is of special interest. Although the function of *TMPRSS2* remains unknown (12, 15), *TMPRSS1*, also known as hepsin, frequently is overexpressed in

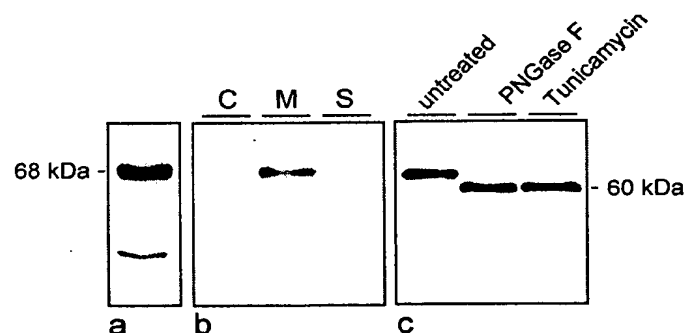


Fig. 4. Western blot analysis of V5-tagged *TMPRSS3* protein. Protein extracts from *TMPRSS3*-pcDNA6/V5/His-transfected HEK-293 cells were resolved in 9% SDS-PAGE and transferred to nitrocellulose membranes. Membranes were immunoblotted with an anti-V5-horseradish peroxidase antibody followed by chemiluminescence detection. *a*, 20 μg of total protein extract. *b*, subcellular localization; C, cytosolic fraction; M, plasma membrane-enriched fraction; S, concentrated culture medium. *c*, analysis of *N*-linked glycosylation of the *TMPRSS3* protein. A shift in molecular mass was detected both after PNGase F treatment of the immunoprecipitated protein and after exposure of the transfected cells to tunicamycin, indicating *N*-glycosylation of the protein.

ovarian tumors and may therefore contribute to the invasive nature or growth capacity of ovarian tumor cells (16). Treatment of hepatoma cells with antihepsin antibodies or specific antisense oligonucleotides confirmed that hepsin plays an essential role in cell growth and maintenance of cell morphology (17). It has also been shown that hepsin can proteolytically activate human coagulation factor VII and thereby contribute to the activation of the coagulation cascade (18).

The correlation of *TMPRSS3* expression with the metastatic potential of the SUIT-2 cell lines is a first indication that this new protease, in the same way as hepsin, may be involved in promoting metastasis formation and tumor invasion. To confirm this hypothesis in functional assays, stably transfected S2-020 cell lines were generated using the *TMPRSS3* cDNA cloned in sense and antisense orientation into the pH β -Apr1-neo vector. Several clones were generated showing variable degrees of *TMPRSS3* sense/antisense mRNA transcription. Two sense and two antisense clones were further characterized by s.c. injections in nude mice, *in vitro* Matrigel invasion assays, and biochemically for their capacity to hydrolyze substrates for trypsin and elastase. No significant differences could be observed between sense and antisense clones in any of the functional assays. There was no difference in tumor size and local invasiveness after s.c. injections, and there was no evidence of metastasis formation after tail vein injection with both sense and antisense cells. Similarly, we failed to show an effect on *in vitro* invasiveness and on proteolytic activity of native and enterokinase-treated lysates for a selection of serine protease substrates. Many factors may be responsible for the failure of *TMPRSS3*-transfected tumor cells to behave differently in these assay, including the necessity for a complex activation mechanism, processes that affect protein folding, or the absence of essential cofactors. Furthermore, although transiently transfected HEK-293 cells showed expression of the V5-tagged recombinant *TMPRSS3* protein, we could not directly demonstrate expression of the protein in the transfected cells because we lacked a specific antibody. In the absence of final experimental proof, we can therefore only hypothesize, based on the structural characteristics and the expression pattern in cancer tissues and in the SUIT-2 subclones, that this new protease has a potential role for tumor progression, metastasis formation, and tumor invasion.

Proteases have an important function in the context of tumor growth, because they can break down the surrounding extracellular matrix components, they can pave the way for spreading tumor cells, and they can release and activate growth and angiogenic factors. Protease activity on the surface of tumor cells is required to allow malignant invasion through surrounding connective tissue, which is an important event in the multistep process of metastasis formation (19). Thus, it is conceivable that *TMPRSS3* may contribute to the invasive and metastatic potential of tumor cells. In this context, cell surface proteases such as *TMPRSS3* may function as an activator of other extracellular proteases or act directly by degrading the extracellular matrix surrounding the tumor cells. Furthermore, *TMPRSS3*, as shown for many other proteases, may participate in the activation of hormones or growth factors by proteolytic cleavage of inactive pro-forms. Because the biochemical events required for the activation of

this novel serine protease are unknown and the specific substrates have not yet been identified, the precise role of *TMPRSS3* in carcinogenesis remains to be elucidated.

Acknowledgments

We thank G. Adler for continual support, U. Lacher for excellent technical assistance, M. A. Hollingsworth for the pH β -Apr1-neo vector, and F. Gansauge and G. Varga for providing human pancreatic tissue samples.

References

- DeClerck, Y. A., and Imren, S. Protease inhibitors: role and potential therapeutic use in human cancer. *Eur. J. Cancer*, **30A**: 2170–2180, 1994.
- Rawlings, N. D., and Barrett, A. J. Families of serine peptidases. *Methods Enzymol.*, **244**: 19–61, 1994.
- Gress, T. M., Wallrapp, C., Frohme, M., Muller-Pillasch, F., Lacher, U., Friess, H., Buchler, M., Adler, G., and Hohelsel, J. D. Identification of genes with specific expression in pancreatic cancer by cDNA representational difference analysis. *Genes Chromosomes Cancer*, **19**: 97–103, 1997.
- Taniguchi, S., Iwanura, T., and Katsuki, T. Correlation between spontaneous metastatic potential and type I collagenolytic activity in a pancreatic cancer cell line (SUIT-2) and sublines. *Clin. Exp. Metastasis*, **10**: 259–266, 1992.
- Gunning, P., Leavitt, J., Muscat, G., Ng, S. Y., and Keddes, L. A human β -actin expression vector system directs high-level accumulation of antisense transcripts. *Proc. Natl. Acad. Sci. USA*, **84**: 4831–4835, 1987.
- Lutz, M. P., Pinon, D. I., Gates, L. K., Shenolikar, S., and Miller, L. J. Control of cholecystokinin receptor dephosphorylation in pancreatic acinar cells. *J. Biol. Chem.*, **268**: 12136–12142, 1993.
- Elbein, A. D. Inhibitors of the biosynthesis and processing of N-linked oligosaccharides. *CRC Crit. Rev. Biochem.*, **16**: 21–49, 1984.
- Dasgupta, A., Takahashi, K., Cutler, M., and Tanabe, K. K. O-Linked glycosylation modifies CD44 adhesion to hyaluronate in colon carcinoma cells. *Biochem. Biophys. Res. Commun.*, **227**: 110–117, 1996.
- Kruger, B., Lerch, M. M., and Tessenow, W. Direct detection of premature protease activation in living pancreatic acinar cells. *Lab. Invest.*, **78**: 763–764, 1998.
- Hartmann, E., Rapoport, T. A., and Lodish, H. F. Predicting the orientation of eukaryotic membrane-spanning proteins. *Proc. Natl. Acad. Sci. USA*, **86**: 5786–5790, 1989.
- Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davic, E. W. A novel trypsin-like serine protease (hepsin) with a putative transmembrane domain expressed by human liver and hepatoma cells. *Biochemistry*, **27**: 1067–1074, 1988.
- Paoloni Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C., and Antonarakis, S. E. Cloning of the *TMPRSS2* gene, which encodes a novel serine protease with transmembrane, LDLRA, and SRCR domains and maps to 21q22.3. *Genomics*, **44**: 309–320, 1997.
- Steitz, T. A., Henderson, R., and Blow, D. M. Structure of crystalline α -chymotrypsin. 3. Crystallographic studies of substrates and inhibitors bound to the active site of α -chymotrypsin. *J. Mol. Biol.*, **46**: 337–348, 1969.
- Anisowicz, A., Sotiropoulou, G., Stenman, G., Mok, S. C., and Sager, R. A novel protease homolog differentially expressed in breast and ovarian cancer. *J. Mol. Med.*, **2**: 624–636, 1996.
- Lin, B., Ferguson, C., White, J. T., Wang, S., Vessella, R., Truc, L. D., Hood, L., and Nelson, P. S. Prostate-localized and androgen-regulated expression of the membrane-bound serine protease *TMPRSS2*. *Cancer Res.*, **59**: 4180–4184, 1999.
- Tanimoto, H., Yan, Y., Clarke, J., Korourian, S., Shigemasa, K., Parmley, T. H., Parham, G. P., and O'Brien, T. J. Hepsin, a cell surface serine protease identified in hepatoma cells, is overexpressed in ovarian cancer. *Cancer Res.*, **57**: 2884–2887, 1997.
- Torres Rosado, A., O'Shea, K. S., Tsuji, A., Chou, S. H., and Kurachi, K. Hepsin, a putative cell-surface serine protease, is required for mammalian cell growth. *Proc. Natl. Acad. Sci. USA*, **90**: 7181–7185, 1993.
- Kazama, Y., Hamamoto, T., Foster, D. C., and Kisiel, W. Hepsin, a putative membrane-associated serine protease, activates human factor VII and initiates a pathway of blood coagulation on the cell surface leading to thrombin formation. *J. Biol. Chem.*, **270**: 66–72, 1995.
- Chen, W. T., Olden, K., Bernard, B. A., and Chu, F. F. Expression of transformation-associated protease(s) that degrade fibronectin at cell contact sites. *J. Cell Biol.*, **98**: 1546–1555, 1984.
- Kyte, J., and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**: 105–132, 1982.

Exhibit 40

MECHANISM OF PROTEIN TRANSLOCATION ACROSS THE ENDOPLASMIC RETICULUM MEMBRANE

Peter Walter and Vishwanath R. Lingappa

Department of Biochemistry and Biophysics and Departments of
Physiology and Medicine, University of California Medical School,
San Francisco, California 94143

CONTENTS	
INTRODUCTION.....	499
HISTORICAL BACKGROUND.....	500
MECHANISM OF TARGETING.....	501
<i>Signal Recognition Particle</i>	502
<i>Signal Sequences</i>	507
<i>SRP Receptor</i>	508
MECHANISM OF TRANSLOCATION.....	509
<i>Machinery</i>	509
<i>Translocation Substrates</i>	510
<i>Posttranslational Translocation in Yeast</i>	511
<i>Posttranslational Translocation of Genetically Engineered Substrates</i>	512
CONCEPTS AND CONTROVERSIES.....	512

INTRODUCTION

In this review we attempt a timely survey of issues concerning protein translocation across the membrane of the endoplasmic reticulum of eukaryotic cells. We focus on recent developments, open questions and current controversies. Due to limited space, this review cannot be and is not

intended to be comprehensive. Where appropriate, reference to more detailed reviews is given in the text.

Eukaryotic cells contain a multiplicity of membrane-delimited compartments. The selective localization of particular proteins provides the basis for each of these compartments to serve various specialized functions. Thus, for example, the mitochondrion is the exclusive residence of enzymes involved in oxidative phosphorylation; similarly, oxidative detoxification takes place exclusively in the endoplasmic reticulum (ER). The proteins that compose, and are contained within, particular membrane systems are kept there by the impermeability of the lipid bilayer to diffusion of proteins across membranes. How then is compartmentalization of newly synthesized proteins achieved, in view of the fact that the cytosol is the common site of synthesis for the majority of proteins, though they are destined for distinct subcellular locations? The term intracellular protein topogenesis has been coined (Blobel 1980) to describe the specialized mechanisms by which newly synthesized proteins selectively overcome the permeability barrier of specific intracellular membranes to achieve their correct subcellular localization. This review addresses the question of how proteins that pass through or reside in the intracisternal space are specifically synthesized on membrane-bound ribosomes and translocated into the ER lumen.

As in the study of other protein translocation events (e.g. across mitochondrial membranes) there are two fundamental issues to resolve regarding transport across the ER membrane: (a) How is the target membrane recognized and distinguished from all other membrane systems? (b) Once it has been targeted, how is the polypeptide chain translocated across the lipid bilayer into the lumen of the organelle?

HISTORICAL BACKGROUND

The work of Palade and coworkers on the secretory pathway (reviewed by Palade 1975) focused attention on ribosomes bound to the rough endoplasmic reticulum as the site of synthesis of secretory proteins. The subsequent demonstration of vectorial discharge of puromycin-released polypeptides into the lumen of isolated rough microsomal vesicles (Redman & Sabatini 1966) suggested that a specialized mechanism was responsible for translocation across the ER membrane: Nascent polypeptides emerged into the lumen of the microsomal vesicles concomitant with their synthesis. These results raised the intriguing question of how the cell could distinguish the mRNAs for secretory proteins from those for cytoplasmic or mitochondrial proteins and selectively translate the former on ER-bound ribosomes.

The signal hypothesis (Blobel & Dobberstein 1975) was proposed to account for these phenomena. Over the last 15 years overwhelming evidence has accumulated from a plethora of experimental systems in favor of this model. As it specifically relates to secretory proteins, the essential tenets of an updated version of this hypothesis (for a recent review see Walter et al 1984) are that: (a) the information for localization of newly synthesized proteins into the lumen of the ER is encoded in a discrete segment of the nascent polypeptide, the signal sequence; (b) this signal sequence interacts with a series of receptors, some of them cytoplasmic, others integral to the ER membrane. Some of these receptors function in targeting the chain to the ER membrane, others function in its actual translocation across that membrane. These latter receptors, together with associated proteins in the ER membrane, constitute the "translocon," a postulated engine able to drive signal sequence-bearing chains across the ER membrane through a proteinaceous pore or channel.

More recently, the concepts of the signal hypothesis have been expanded to describe a general framework for intracellular protein topogenesis (Blobel 1980). According to this model, "topogenic sequences" within discrete segments of targeted proteins are decoded by specific receptors, either during (cotranslational) or shortly after (posttranslational) their biosynthesis. The specificity of such signal sequence-receptor interactions targets the proteins to the correct intracellular membranes where they are fed into translocons that move them across the hydrophobic core of the lipid bilayer. Similarly, it has been proposed that another class of topogenic sequences—termed stop-transfer sequences—interacts with the translocon to arrest further transport and thereby achieve an asymmetric transmembrane orientation of integral membrane proteins. Thus many of the concepts developed in this review for soluble ectoplasmic proteins are directly applicable to the problem of integration of transmembrane proteins. Recent developments reviewed below suggest that translocons in different intracellular membrane systems may function more similarly than previously thought.

MECHANISM OF TARGETING

With the availability of *in vitro* systems that faithfully reproduce the translocation of nascent proteins [secretory proteins (Blobel & Dobberstein 1975), lysosomal proteins (Erickson et al 1983), and certain classes of integral membrane proteins (Katz et al 1977)], it became feasible to investigate the molecular requirements for protein translocation across the ER membrane. So far, two components, the signal recognition particle



(SRP) and the SRP receptor, have been purified and shown to function in the targeting events preceding the actual translocation event.

Signal Recognition Particle

SRP is an 11S small cytoplasmic ribonucleoprotein (Walter & Blobel 1982). In our current view, SRP functions as an adapter between the protein synthetic machinery in the cytoplasm and the protein translocation machinery in the ER membrane.

STRUCTURE OF SRP SRP was first recognized by its ability to restore the translocation activity of salt-extracted microsomes in vitro (Warren & Dobberstein 1978). It was purified to homogeneity from a salt extract of canine pancreatic microsomal vesicles using this activity as an assay (Walter & Blobel 1980). SRP consists of a small (300 nucleotide) 7SL RNA (Walter & Blobel 1982) and six nonidentical polypeptide chains organized into four SRP proteins. These proteins are two monomers, a 19-kDa polypeptide and a 54-kDa polypeptide, and two heterodimers, one composed of a 9-kDa and a 14-kDa polypeptide, and the other comprised of a 68-kDa and a 72-kDa polypeptide (Siegel & Walter 1985). When SRP is disassembled under nondenaturing conditions, the RNA and the protein fractions are inactive by themselves, but together they can readily be reconstituted into an active particle (Walter & Blobel 1983; Siegel & Walter 1985).

Recent studies revealed that different assayable functions of SRP in the targeting process can be assigned to specific structural domains of the particle. These separable functions include the recognition of signal sequences and the ability of SRP to arrest specifically the translation of nascent signal sequence-bearing proteins (Siegel & Walter 1986b). These domains are schematically indicated in Figure 1 superimposed on the secondary structure of 7SL RNA. This model is supported by recent evidence demonstrating that SRP is a rod-shaped, elongated structure (Andrews et al 1985) and that the RNAs—visualized directly by electron spectroscopic imaging—span the entire length of the particle (D. W. Andrews et al, submitted for publication).

SIGNAL RECOGNITION Once SRP had been purified to homogeneity it became possible to study its activity in greater detail. Results of experiments testing both the effects of SRP on the translation of secretory proteins and its binding properties with various components in the translation-translocation system have led to the model of the SRP cycle shown in Figure 2.

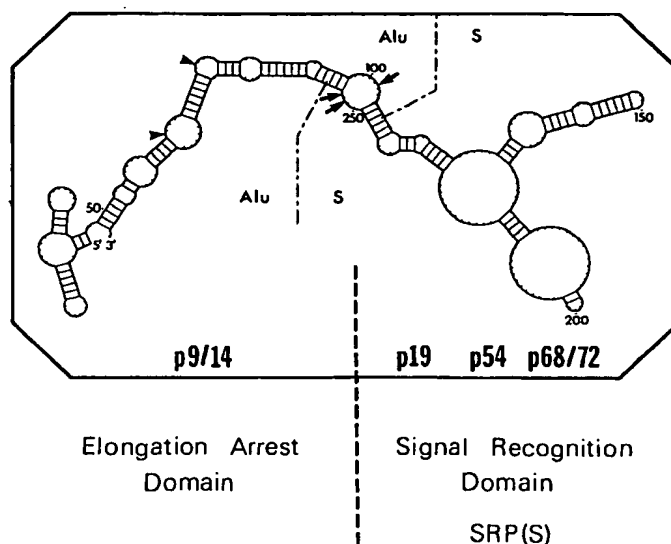
In brief, SRP is thought to bind in a signal-sequence-independent

manner with relatively low affinity to biosynthetically inactive ribosomes (Figure 2a, b) (Walter et al 1981). Upon emergence of a signal sequence as part of the nascent polypeptide chain, the affinity of SRP for the ribosome increases (Figure 2c); in the case of preprolactin synthesized on wheat germ ribosomes this increase amounts to three to four orders of magnitude. The SRP-ribosome-nascent chain complex is then targeted to the membrane of the ER via a direct interaction of SRP with the SRP receptor (Walter & Blobel 1981b), an integral membrane protein that is restricted in its subcellular localization to this membrane system (Hortsch et al 1985). At this point SRP and the SRP receptor detach from the ribosome and can reenter the cycle, i.e. both molecules are thought to act catalytically in the targeting process. The ribosome-nascent chain complex engages in a functional ribosome membrane junction, and the translocation of the nascent polypeptide proceeds (see below). (For a more detailed description of the SRP cycle see Walter et al 1984.)

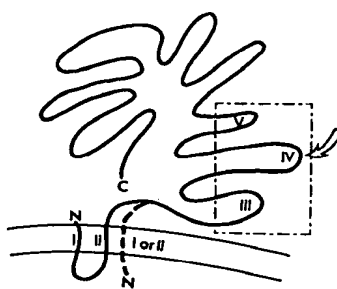
ELONGATION ARREST When SRP is included in *in vitro* translation systems in the absence of microsomal membranes, it blocks protein synthesis concomitant with the increase in its affinity for the ribosome just after the signal peptide becomes exposed outside the large ribosomal subunit (Walter & Blobel 1981b; Meyer et al 1982a). In some cases a discretely sized protein fragment that corresponds to the elongation-arrested secretory protein can be detected by gel electrophoresis; in other cases the arrested forms appear as a broader smear on gels, which indicates that SRP can recognize signal sequences and arrest elongation within a certain range of chain lengths. It is also observed that some nascent polypeptides are arrested, while others transiently pause in chain growth (P. Walter, unpublished results). Therefore, in these latter cases arrest is often difficult to detect (Meyer 1985). Interestingly, while elongation arrest has been demonstrated as a kinetic delay of elongation in translation systems reconstituted from mammalian components (K. Matlack & P. Walter, unpublished results), the same effect is more pronounced (as a strict blockage of elongation) when signal-bearing proteins are translated in a heterologous wheat germ system. Thus while the general phenomenon of arrested elongation is ubiquitous, different *in vitro* systems reflect it to a different degree. Therefore it remains to be established whether SRP acts *in vivo* as a strict "on-off" switch or functions as a more graded rate-controlling factor.

Two distinct biochemical approaches were employed to map the elongation-arrest function to a separate and separable domain of SRP. One functional domain was shown to consist of the 9/14-kDa SRP proteins and those 7SL RNA sequences that are homologous to repetitive Alu DNA (see Figure 1, *left*). One experimental approach employed single omission

experiments in which SRPs were reconstituted from fractionated and purified protein and RNA components (Siegel & Walter 1985). A second approach involved the preparation of a subparticle obtained after nucleolytic dissection of SRP (Siegel & Walter 1986). These perturbed SRPs lacking the elongation-arrest domain are still active in signal recognition and targeting; therefore, elongation arrest *cannot* be a prerequisite for protein translocation across the membrane. In the absence of elongation arrest, however, most signal-bearing nascent proteins lose their ability to



(a)



(b)

be translocated if elongation proceeds beyond a critical point in the absence of membranes. Thus elongation arrest seems to maintain the nascent chain in a translocation-competent state by preventing (or delaying) its further elongation into the cytoplasmic space and thereby adds to the fidelity of the reaction. The particular length range in which a nascent protein remains translocation competent may vary for different proteins (see below).

Since SRP contains an RNA as a structural component, it is tempting to speculate that this RNA engages in base-pairing interactions with other nucleic acids during the SRP's functional cycle. The RNA components in the translational apparatus are likely candidates for participants in such interactions (Walter & Blobel 1982; Zwieb 1985). However, there is at present no direct evidence for such interactions. A possible mechanism for elongation arrest could involve the binding of 7SL RNA to the A-site on the ribosome, thus preventing the next amino acyl tRNA from binding. Indeed, the secondary structure of 7SL RNA in the elongation-arrest

Figure 1 Domain structure of SRP (*left*) and the SRP receptor (*right*). (*a*) (From Siegel & Walter 1986a): SRP is composed of two separable domains. A possible phylogenetically conserved secondary structure for 7SL RNA is shown (Siegel & Walter 1986a). Similar secondary structures have been proposed by Gundelfinger et al (1984), E. Ullu (personal communication), and Zwieb (1985). Connecting lines between the RNA strands indicate base pairs; G-U pairs are included. (For an extensive description of SRP structure see Siegel & Walter 1986b.) Micrococcal nuclease cleaves the particle at the point indicated by arrows, removing the elongation-arresting domain. Additional cuts mapped by Gundelfinger et al (1983) are indicated by arrowheads. The elongation-arresting domain includes both ends of the RNA (labeled 5' and 3') and is comprised of sequences that are homologous to the repetitive Alu DNA sequence family. Evolutionary considerations suggest that 7SL RNA is the parent molecule for repetitive Alu DNA (Ullu & Tschudi 1985). The thin dashed lines indicate the boundaries of homology between 7SL RNA and an Alu consensus sequence. The elongation-arresting domain also contains the 9/14-kDa SRP protein. The other domain, termed SRP(S), retains signal recognition and translocation promoting function and is comprised of the middle portion of 7SL RNA (the S-segment) and the remaining three SRP proteins. As mentioned in the text, the 54-kDa SRP protein can be selectively cross-linked to signal peptides and may therefore provide the signal binding pocket. (*b*) (From Lauffer et al 1985): A model of the disposition of the SRP receptor α -subunit in the membrane of the ER is shown. Putative structural and functional features as deduced from the primary sequence (Lauffer et al 1985) are indicated. Regions I and II are putative membrane-spanning regions; whether both of them or either one alone functions as the membrane anchor of the receptor or if additional hydrophobic regions are contributed by the β -subunit is presently not known. Regions III-V contain the charge clusters described in the text. The boxed domain contains regions strongly resembling RNA binding proteins; their presence suggests that the SRP-SRP receptor interaction may include binding of 7SL RNA to this domain. The arrow indicates the position of the protease-sensitive site. Cleavage of the receptor at this position results in the release of the 52-kDa cytoplasmic fragment. This fragment does not have two properties of the intact receptor: the binding affinity for SRP and the ability to release elongation arrest (Lauffer et al 1985; Gilmore et al 1982a).

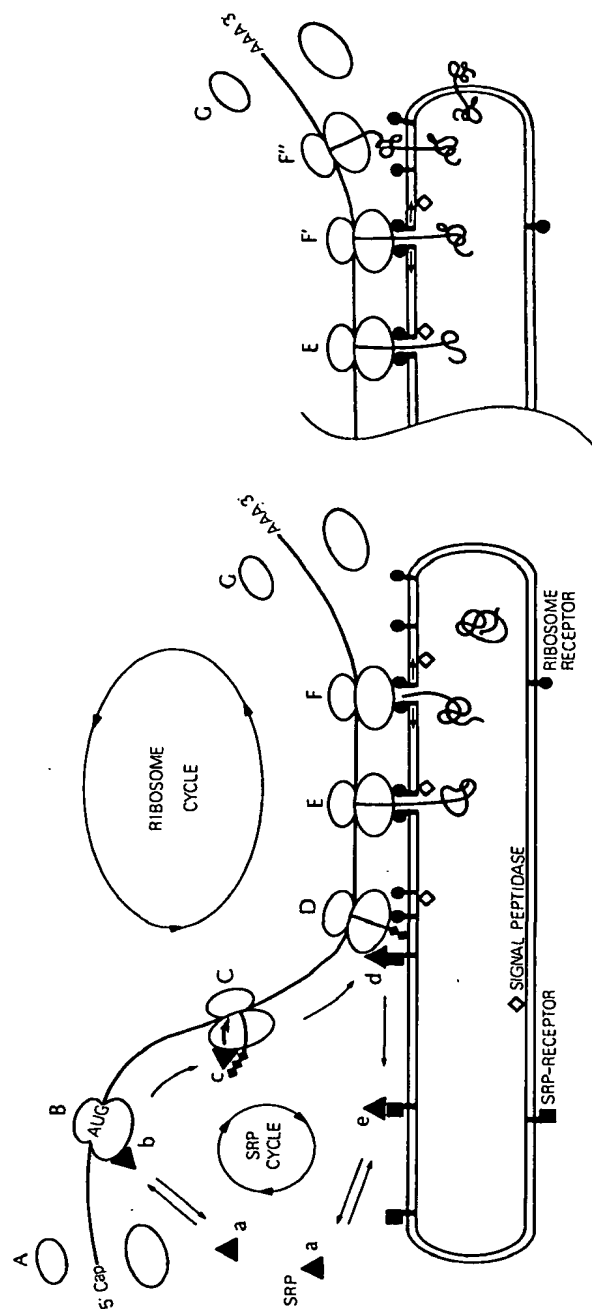


Figure 2 Model (from Walter et al 1984) for protein translocation across the ER membrane for soluble intracisternal proteins (*left*) and integral membrane proteins that possess a structural domain on the intracisternal face of the membrane (*right*). The key features of the model are outlined in the text. (For a more extensive description see Walter et al 1984.)



domain of SRP resembles that of a tRNA that is missing the anticodon stem. In addition, the physical dimensions of SRP would easily allow the particle to bridge the distance between the nascent chain exit site on the ribosome (where the signal sequence emerges) and the peptidyl transferase activity known to be located between the two ribosomal subunits (Andrews et al 1985).

Signal Sequences

What constitutes the essential features of a signal sequence and how such sequences are recognized by SRP remain unsolved problems. Signal sequences show no recognizable primary sequence homology, and a recent compilation shows that sequence variation can be rather extreme (von Heijne 1985). Yet studies on a variety of systems both in vivo and in vitro demonstrate conservation of signal sequence function over the widest evolutionary distances (Muller et al 1982). As a consequence we are still not able to predict with confidence which regions in proteins might function as internal signal sequences. Nevertheless, internal signal sequences have been demonstrated unequivocally (Bos et al 1984). Moreover, cleavage by signal peptidase is not required for translocation (Palmiter et al 1978).

One of the few characteristic features of signal sequences is a variable stretch of hydrophobic amino acids in the core of the sequence. Point mutations in the hydrophobic core in bacterial signal sequences have been shown to abolish function (Lee & Beckwith 1986, this volume). Based on the hydrophobicity of these regions and on evidence from biophysical studies with synthetic signal peptides (reviewed by Briggs & Gierasch 1986), it has been suggested that these sequences act as amphiphiles that are integrated into and possibly perturb lipid bilayers. There is, however, still no evidence that the general mechanism for translocation involves a direct interaction of signal sequences with the hydrophobic core of the lipid bilayer. Indeed, several lines of evidence suggest direct interactions of signal sequences with *proteins*.

The clearest evidence for such interactions involve SRP. Since SRP is a soluble ribonucleoprotein, its interactions with signal sequences can be studied in the absence of membranes by measuring binding or by observing the SRP-mediated modulation of protein synthesis. For example, when signal sequences that are rich in leucine are translated in the presence of the amino acid analog β -hydroxy-leucine, SRP signal recognition is abolished (Walter et al 1981; Walter & Blobel 1981b). This demonstrates that SRP directly recognizes features in the nascent chain. Moreover, the finding conclusively rules out the possibility that sequences in the mRNA alone are responsible for the observed effect. (After the discovery of an RNA component in SRP the latter notion was considered attractive



because of the possibility of recognition via putative base-pairing interactions.) Direct proof of an SRP-signal sequence interaction was recently provided by cross-linking experiments. Two groups independently showed that a photoactivable cross-linking reagent was selectively incorporated into the amino-terminal region of the signal peptide for nascent preprolactin. Each group found that the signal peptide is in *direct* contact with the 54-kDa SRP protein (Kurzchalia et al 1986; Krieg et al 1986).

SRP Receptor

Using the same in vitro protein translocation assays that led to the purification of SRP, two distinct approaches were taken to identify the corresponding *membrane* components involved in targeting of signal sequence-bearing nascent chains to the ER membrane. These approaches eventually led to the discovery and purification of the SRP receptor, the first membrane protein proven to play a vital role in this process.

One of these approaches was based on the early observation that proteolysis of microsomal membranes completely abolishes their protein translocation activity but that, most importantly, the activity can be restored by addition to an extract prepared by limited proteolysis of the original microsomal membrane fraction (Walter et al 1979; Meyer & Dobberstein 1980a). This proteolytic dissection and functional reconstitution provided the assay for the purification of the protease-solubilized component. The activity was purified as a basic 52-kDa protein (apparent mobility on SDS PAGE is 60 kDa) (Meyer & Dobberstein 1980b), which was subsequently demonstrated (by immunological techniques) to be a proteolytic fragment derived from a 69-kDa integral membrane protein (apparent mobility 72 kDa) restricted in its subcellular localization to the endoplasmic reticulum (Meyer et al 1982b).

The second approach took advantage of the observations that, when assayed in the absence of microsomal membranes, SRP causes a site-specific elongation arrest in the synthesis of presecretory proteins and that microsomal membranes contain an activity that releases the elongation arrest. Based on these observations, the elongation-arrest-releasing activity was predicted to reside in a membrane protein termed the SRP receptor (Walter & Blobel 1981b) [subsequently named the docking protein (Meyer et al 1982a)]. Fractionation of a detergent extract of microsomal membranes employing affinity chromatography on SRP-Sepharose as a key step allowed purification of the SRP receptor. The purified fraction contained a predominant 69-kDa membrane protein and the arrest-releasing activity. Using both immunological and peptide-mapping techniques, the SRP receptor was shown to be identical to the membrane protein identified via the proteolytic dissection methods described above (Gilmore et al 1982a,b).

Recently, the primary structure of the 69-kDa SRP receptor protein was determined from its cognate cloned cDNA, and its relationship to the cytoplasmic SRP receptor fragment was determined (Lauffer et al 1985). This fragment was shown to begin with residue 152 of the intact protein. Thus, it is sequences within the 151 amino acids at the amino terminal that anchor the SRP receptor in the lipid bilayer. Two distinctly hydrophobic regions have been identified that constitute putative α -helical trans-membrane segments. Since either of these segments would position a positively charged amino acid in the hydrophobic core of the lipid bilayer, the receptor probably interacts with other integral membrane proteins that neutralize these charges. Recent evidence suggests the existence of proteins that can be copurified with the 69-kDa SRP receptor protein or isolated by affinity techniques. In particular, an ER membrane protein with an apparent molecular weight of 30 kDa was found by a variety of techniques to be tightly associated with the 69-kDa protein (Tajima et al 1986). Thus the SRP receptor appears to be a hetero-dimeric protein that in addition to the 69-kDa polypeptide (the SRP receptor α -subunit) contains a second 30-kDa subunit (β -subunit). Carboxy-terminal to the putative trans-membrane regions in the α -subunit is an unusually hydrophilic domain. In particular, unusually large clusters of charged amino acids are found surrounding the site of proteolytic cleavage that severs the 52-kDa cytoplasmic domain (see Figure 1, *right*). This domain of the SRP receptor strongly resembles nucleic acid binding proteins, which suggests that the receptor may transiently interact directly with the 7SL RNA in SRP and that the SRP-SRP receptor affinity could be mediated, at least in part, by a protein-nucleic acid interaction.

The SRP receptor is unlikely to be part of the translocon itself, because the receptor is present in the ER membrane in substoichiometric amounts with respect to membrane-bound ribosomes. Thus it was suggested that the SRP receptor functions "catalytically" and is recycled once correct targeting of the ribosome has been achieved (Gilmore & Blobel 1983). There is also evidence for an additional activity that is distinct from SRP and the SRP receptor and may interact with the targeted signal sequence and act as a secondary signal receptor(s) in the ER membrane (Gilmore & Blobel 1985; Prehn et al 1980). However, a protein serving this function has not yet been identified.

MECHANISM OF TRANSLOCATION

Machinery

Cell-free systems provided a detailed molecular description of the targeting machinery, but have yet to allow insights into the molecular details of the



translocation process. In part this difficulty results from the apparent obligate coupling of translocation and translation: Transport across the ER membrane takes place cotranslationally; completed precursors are not detectable in vivo in the cytoplasm. In cell-free systems translocation proceeds only during a limited time and under the fastidious conditions required for the synthesis of the very molecule whose translocation is being studied. As a result, although several specific polypeptides have been implicated as functional components of the translocon, the direct role of any of these proteins remains to be demonstrated. For example, two integral membrane proteins, termed ribophorins, have been suggested to act as ribosome receptors (Kreibich et al 1978); the recent purification of signal peptidase, a relatively abundant complex of six polypeptides, suggests that these proteins are involved in other functions besides signal cleavage (Evans et al 1986).

Translocation Substrates

Although we know little about the actual machinery involved, insight into certain aspects of the mechanism of translocation has recently been obtained by approaches involving manipulation of the translocation substrates. For example, expression of engineered cDNAs encoding fusion proteins in transcription-linked translation systems demonstrated that a signal sequence was sufficient to direct translocation of normally cytoplasmic globin, both in vitro (Lingappa et al 1984) and in vivo (K. Simon et al, submitted for publication). Thus, the specific information for translocation was contained within the signal sequence and not the "passenger" protein.

A more complex version of these experiments raised interesting questions as to the mechanism of translocation (Perara & Lingappa 1985). The DNA sequence coding for globin, normally a cytosolic protein, was fused with the 5' end of the DNA sequence for preprolactin, a secretory protein that has an amino-terminal signal sequence. This fusion protein thus contained the preprolactin signal sequence at an internal position, 117 amino acids from the initiator methionine. When expressed in a transcription-linked translation system, this internal signal sequence was not only cleaved by signal peptidase, but directed the translocation of both flanking protein domains. Surprisingly, carbonate extraction demonstrated that neither the globin domain with the signal sequence attached at its carboxy terminus nor the prolactin domain were integrated into the membrane. Instead, both resided in the vesicle lumen either free or bound to proteins. This result suggests that signal sequences are not buried in the bilayer directly but perform their function by interacting with a protein-

aceous machinery in the membrane. Moreover, translocation of the globin domain by a subsequently emerging signal sequence suggests that the energy used for the globin domain's synthesis is not required for its translocation. Thus the commonly observed coupling of translocation and translation may not be an obligate requirement for transport across the ER membrane.

The notion that the translocation machinery can function independently of protein synthesis has now received direct support from different experimental systems.

Posttranslational Translocation in Yeast

Recently, *in vitro* translation-translocation systems from the yeast *Saccharomyces cerevisiae* have been established (Hansen et al 1986; Waters & Blobel 1986; Rothblatt & Meyer 1986). The precursor to the yeast pheromone α -factor has been used as a model secretory protein. Contrary to all expectations, this precursor, an ~ 18.5 kDa protein, is translocated across yeast ER membranes posttranslationally, i.e. after it has been completely synthesized and has been released from ribosomes. Prepro- α -factor has no particularly hydrophobic or amphipathic stretches in its primary sequence (other than a typical signal sequence), making it unlikely that its posttranslational translocation is due to some passive partitioning of the protein across the lipid bilayer. Furthermore, the posttranslational translocation reaction is ATP-dependent and requires protein elements both in the membrane and the soluble fraction. Whether these protein components are related in any way to the putative yeast SRP and SRP receptor analogs remains to be established by biochemical analysis. It is clear from these data, however, that translocation of prepro- α -factor does not require coupling to protein synthesis. Therefore, the translocon can, in principle, accept its substrate posttranslationally and in the absence of the ribosome.

It should be kept in mind that the posttranslational translocation of prepro- α -factor was observed *in vitro* in a system artificially depleted of ER membranes during synthesis. This finding does not prove that prepro- α -factor ever crosses the ER membrane posttranslationally *in vivo*, where ER membranes are always present during translation. Rather, the actual degree of coupling of translocation and protein synthesis will depend on the relative rates of the respective processes. If targeting and translocation are fast with respect to protein elongation, a strictly vectorial cotranslational translocation mode will result, as appears to be the rule in mammalian cells *in vivo* (Bergman & Kuehl 1979; Glabe et al 1980).



Posttranslational Translocation of Genetically Engineered Substrates

Similar findings also emerged from the use of engineered clones in mammalian cell-free translation systems (Perara et al 1986; Mueckler & Lodish 1986). Using a procedure that generates a truncated mRNA lacking a termination codon, secretory polypeptide chains could be synthesized and presented to membranes in the absence of further chain elongation while still held by the ribosome that effects their synthesis. It was demonstrated that such chains could be translocated and that nucleotide triphosphates were required as the energy source for this process. In contrast to the situation in the yeast system described above, in most of these cases translocation could be abolished by releasing the nascent chain from the ribosome by artificial termination with the amino acyl tRNA analog puromycin. As expected, translocation was abolished by deletion of the coding region for the signal sequence. In some cases, however, it was also found that some short chains could translocate in a ribosome-independent condition analogous to that found for prepro- α -factor in the yeast system (E. Perara & V. R. Lingappa, submitted for publication). Thus it appears that, at least for the proteins investigated, polypeptide chain growth proceeds through stages in which translocation competence is a property of the chain itself or is maintained by interaction with the ribosome (see Figure 3).

These results show cotranslational translocation in a new light: The role of the membrane-bound ribosome is not to extrude or push the chain through the bilayer as suggested by some observers (Wickner & Lodish 1985). Rather, translocation is catalyzed by an energy-consuming protein engine in the ER membrane, and the ribosome acts, in most but not all cases, as a ligand that maintains the translocation competence of the nascent chain.

CONCEPTS AND CONTROVERSIES

We have surveyed the development of ideas on the problem of translocation of newly synthesized proteins across the ER membrane. Initially, attention was focused on the coupling of translocation to translation, a feature unique to translocation across the ER membrane. This has given way to the realization that obligate coupling to translation is not a prerequisite for translocation and that transport across membranes of a variety of organelles may share common features. These include the involvement of a targeting receptor to discriminate among proteins intended for different destinations, a translocon that somehow transports

the targeted protein across the bilayer, and a requirement for energy (derived from hydrolysis of nucleoside triphosphates or from an electrochemical gradient) to drive translocation. The recognition of these steps has resulted from the study of diverse proteins in a variety of organisms and from the study of "artifacts" generated in vitro, i.e. biochemically or genetically altered translocation machinery (Siegel & Walter 1986b) and substrates (Perara & Lingappa 1985), whose aberrant behavior has provided insight into fundamental details of the targeting and translocation problem. Even as new questions emerge, many old ones (e.g. the molecular nature of the signal sequence-receptor interaction) remain unanswered.

Other questions must now be reformulated. For example, in spite of the recent demonstration that the translocon in the ER membranes can, in principle, accept translocation substrates posttranslationally, translocation most likely occurs cotranslationally in vivo. The observation that most posttranslational translocation across the ER membrane appears to be ribosome dependent in vitro supports this notion. As described earlier, ribosome-independent and ribosome-dependent modes of posttranslational translocation across the ER membrane probably reflect the requirements for maintenance of the "translocation competent state" of the nascent chain (see Figure 3). Loss of translocation competence may be due to folding (aberrant or normal) or oligomerization of the protein, or entanglement of the signal sequence with the rest of the chain such that the resulting structure can no longer functionally interact with either the targeting or translocation machinery. A few proteins (such as yeast prepro- α -factor) retain translocation competence even as free, completed polypeptides. For most proteins, however, translocation competence is restricted to a generally narrow range of chain lengths. This range can be extended if the polypeptide is targeted to the membrane while still attached to the ribosome. However, eventually most proteins reach a point in chain elongation where translocation competence is no longer maintained, even when the protein is associated with the ribosome. One of the roles of the SRP-induced elongation arrest may therefore be to extend the effective range of translocation competence for the nascent polypeptide chains.

Previously, the nascent chain was thought to be vectorially translocated across the membrane as it emerged from the ribosome; the finding of posttranslational translocation raises the possibility that the translocon may be sufficiently pliable to accept (partially) folded domains rather than exclusively linear polypeptide chains. Alternatively, the translocon may effect unfolding of such domains prior to translocation. In either case the molecular environment traversed by the protein as it passes through the bilayer remains to be investigated. The finding that translocation is driven by nucleoside triphosphate hydrolysis is a direct demonstration of a protein

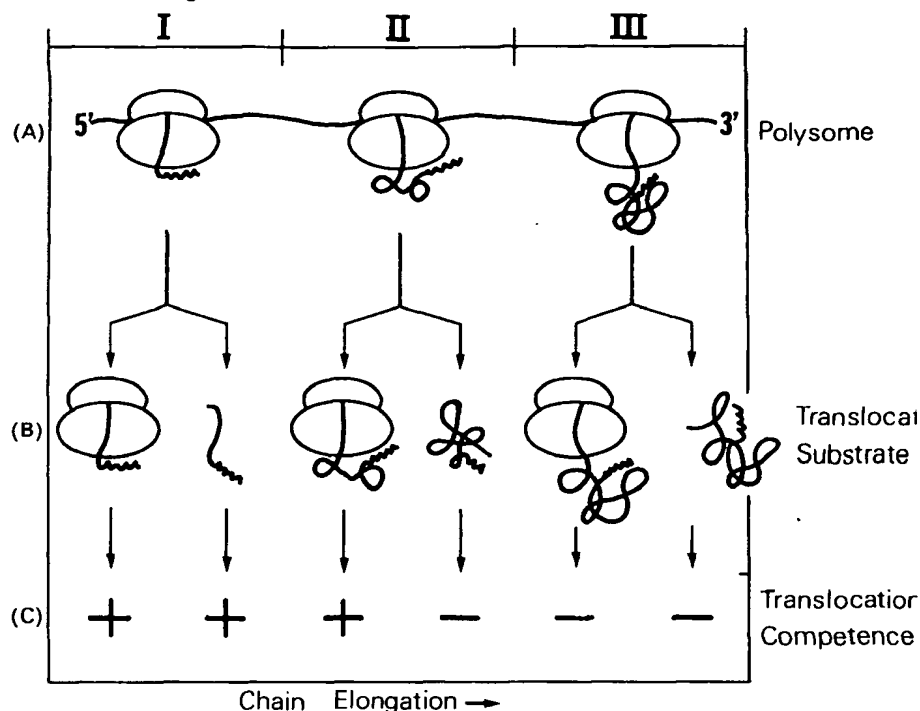


Figure 3 Ribosome dependence of translocation competence. This figure depicts the natural history of the relationship of chain growth (*A*) to translocation competence (*C*). The ribosome dependence of posttranslational translocation was assayed for various lengths of polypeptide synthesized. Progressively shorter polypeptides were synthesized by translating mRNA transcripts in vitro that were progressively truncated at their 3' end and therefore lacked termination codons (Perara et al 1986; E. Perara & V. R. Lingappa, manuscript in preparation). Ribosomes that have reached the 3' end of such a truncated mRNA appear unable to release the newly synthesized polypeptide. Release can be artificially achieved by treatment with puromycin. Such translocation substrates, either with or without release from the ribosomes (as indicated in *B*), can be assayed for translocation competence upon presentation to a microsomal membrane preparation in the presence of nucleoside triphosphate to supply energy. In this assay the ribosome dependence or independence of the translocation competence is reflected in the ability or inability of puromycin pretreatment to abolish translocation by releasing the chain from the ribosome (see right arms of branched arrows). (*A*) depicts three ribosomes on a polysome at various stages (I, II, and III) during the synthesis of a hypothetical secretory polypeptide chain. In (*C*) translocatin competence as assayed posttranslationally (see above) is indicated (+). At stage I, the nascent chain is translocation competent, and this competence is independent of the presence of the ribosome, as experimentally demonstrated. As chain growth proceeds, the polypeptide enters stage II where its translocation competence requires the ribosome. Finally, late in chain growth (stage III) the chain is no longer competent to interact with receptors and other proteins involved in translocation. Whether loss of translocation competence in stage III involves a loss of targeting function or loss of a productive interaction with the translocon remains to be determined. It is not known whether SRP is required for posttranslational translocation in either case.



engine in the membrane and rules out a spontaneous process previously suggested (Wickner 1979; Engelman & Steitz 1980). It remains to be established how the energy of hydrolysis is used by the translocon.

Old controversies regarding co- versus posttranslational translocation appear to be resolved. In retrospect it could be concluded that many prokaryotic proteins (targeted to the plasma membrane) do not require ribosomes to maintain their translocation competence. This also appears to be the case for all proteins (so far studied) that are translocated across the peroxisomal membrane and the mitochondrial and chloroplast envelopes. The most challenging problems for future research now include the further fractionation and purification of all the essential, as well as modulatory, components of the targeting and translocation machinery. This should ultimately allow their reconstitution in *in vitro* systems for the mechanistic analysis of their functions. Finally, our goal must be the understanding of how these components function *in vivo*. This should include elucidation of the regulatory or homeostatic mechanisms involved in harnessing such a remarkable set of protein machines as the translocons.

ACKNOWLEDGMENTS

We wish to thank David Andrews, Patricia Hoben, and Leander Lauffer for many helpful comments on the manuscript. This work was supported by NIH grants GM-32384 to PW and GM-31626 to VRL. PW is a recipient of support from the Chicago Community Trust/Searle Scholars Program.

Literature Cited

- Andrews, D. W., Walter, P., Ottensmeyer, F. P. 1985. *Proc. Natl. Acad. Sci. USA* 82: 785-89
- Bergman, L. W., Kuehl, L. M. 1979. *J. Biol. Chem.* 254: 8869-76
- Blobel, G. 1980. *Proc. Natl. Acad. Sci. USA* 77: 1496-1500
- Blobel, G., Dobberstein, B. 1975. *J. Cell Biol.* 67: 835-51
- Bos, T. J., Davis, A. R., Nayak, D. P. 1984. *Proc. Natl. Acad. Sci. USA* 81: 2337-41
- Briggs, M. S., Gierasch, L. M. 1986. *Adv. Protein Chem.* 38: In press
- Engelman, D. M., Steitz, T. A. 1981. *Cell* 23: 411-22
- Erickson, A. H., Walter, P., Blobel, G. 1983. *Biochem. Biophys. Res. Commun.* 115: 275-80
- Evans, E., Gilmore, R., Blobel, G. 1986. *Proc. Natl. Acad. Sci. USA* 83: 581-85
- Gilmore, R., Blobel, G., Walter, P. 1982a. *J. Cell Biol.* 95: 463-69
- Gilmore, R., Blobel, G. 1983. *Cell* 35: 677-85
- Gilmore, R., Blobel, G. 1985. *Cell* 42: 497-505
- Gilmore, R., Walter, P., Blobel, G. 1982b. *J. Cell Biol.* 95: 470-77
- Glabe, C. G., Hanover, J. A., Lennarz, W. J. 1980. *J. Biol. Chem.* 255: 9236-41
- Gundelfinger, E. D., Carlo, M. D., Zopf, D., Melli, M. 1984. *EMBO J.* 3: 2325-32
- Gundelfinger, E. D., Krause, E., Melli, M., Dobberstein, B. 1983. *Nucleic Acids Res.* 11: 7363-73
- Hansen, W. B., Garcia, P. D., Walter, P. 1986. *Cell* 45: 397-406
- Hortsch, M., Griffiths, G., Meyer, D. I. 1985. *Eur. J. Cell Biol.* 38: 271-79
- Katz, F. N., Rothman, J. E., Lingappa, V. R., Blobel, G., Lodish, H. F. 1977. *Proc. Natl. Acad. Sci. USA* 74: 3278-82
- Kreibich, G., Freenstein, C. M., Pereyra, B. N., Ulrich, B. L., Sabatini, D. D. 1978. *J. Cell Biol.* 77: 488-506
- Krieg, U., Walter, P., Johnson, A. 1986. *Proc. Natl. Acad. Sci. USA*. In press
- Kurzchalia, T. V., Wiedmann, M., Gir-



516 WALTER & LINGAPPA

- shovich, A. S., Bochkareva, E. S., Bielka, H., Rapoport, T. A. 1986. *Nature* 320: 634-36
- Lauffer, L., Garcia, P. D., Harkins, R. N., Coussens, L., Ullrich, A., Walter, P. 1985. *Nature* 318: 334-38
- Lee, C., Beckwith, J. 1986. *Ann. Rev. Cell Biol.* 2: 315-36
- Lingappa, V. R., Chaider, J., Yost, C. S., Hedgpeth, J. 1984. *Proc. Natl. Acad. Sci. USA* 81: 456-60
- Meyer, D. I. 1985. *EMBO J.* 4: 2031-33
- Meyer, D. I., Dobberstein, B. 1980a. *J. Cell Biol.* 87: 498-502
- Meyer, D. I., Dobberstein, B. 1980b. *J. Cell Biol.* 87: 503-8
- Meyer, D. I., Krause, E., Dobberstein, B. 1982a. *Nature* 297: 647-50
- Meyer, D. I., Louvard, D., Dobberstein, B. 1982b. *J. Cell Biol.* 92: 579-83
- Mueckler, M., Lodish, H. F. 1986. *Cell* 44: 629-37
- Muller, M., Ibrahim, I., Chang, C. N., Walter, P., Blobel, G. 1982. *J. Biol. Chem.* 257: 11860-63
- Palade, G. 1975. *Science* 189: 347-58
- Palmiter, R. D., Gagnon, J., Walsh, K. A. 1978. *Proc. Natl. Acad. Sci. USA* 75: 94-98
- Perara, E., Lingappa, V. R. 1985. *J. Cell Biol.* 101: 2292-2301
- Perara, E., Rothman, R. E., Lingappa, V. R. 1986. *Science* 232: 348-52
- Prehn, S., Nurnberg, P., Rapoport, T. A. 1980. *Eur. J. Biochem.* 107: 185-95
- Redman, C. M., Sabatini, D. D. 1966. *Proc. Natl. Acad. Sci. USA* 56: 608-15
- Rothblatt, M., Meyer, D. I. 1986. *Cell* 44: 619-28
- Siegel, V., Walter, P. 1985. *J. Cell Biol.* 100: 1913-21
- Siegel, V., Walter, P. 1986a. *Nature* 320: 81-84
- Siegel, V., Walter, P. 1986b. In *Genetic Engineering*, Vol. 8, pp. 179-94, ed. J. K. Setlow. New York: Plenum
- Tajima, S., Lauffer, L., Rath, V., Walter, P. 1986. *J. Cell Biol.* 103: In press
- Ullu, E., Tschudi, C. 1985. *Nature* 312: 171-72
- von Heijne, G. 1985. *J. Mol. Biol.* 184: 99-105
- Walter, P., Blobel, G. 1980. *Proc. Natl. Acad. Sci. USA* 77: 7112-16
- Walter, P., Blobel, G. 1981a. *J. Cell Biol.* 91: 551-56
- Walter, P., Blobel, G. 1981b. *J. Cell Biol.* 91: 557-61
- Walter, P., Blobel, G. 1982. *Nature* 299: 691-98
- Walter, P., Blobel, G. 1983. *Cell* 34: 525-33
- Walter, P., Gilmore, R., Blobel, G. 1984. *Cell* 38: 5-8
- Walter, P., Ibrahim, I., Blobel, G. 1981. *J. Cell Biol.* 91: 545-50
- Walter, P., Jackson, R. C., Marcus, M. M., Lingappa, V. R., Blobel, G. 1979. *Proc. Natl. Acad. Sci. USA* 76: 1796-99
- Warren, G., Dobberstein, B. 1978. *Nature* 273: 569-71
- Waters, G., Blobel, G. 1986. *J. Cell Biol.* 102: 1543-50
- Wickner, W. T. 1979. *Ann. Rev. Biochem.* 48: 23-45
- Wickner, W. T., Lodish, H. F. 1985. *Science* 230: 400-7
- Zwieb, C. 1985. *Nucleic Acids Res.* 13: 6105-24



CONTENTS

ACTIVATION OF SEA URCHIN GAMETES, <i>James S. Trimmer and Victor D. Vacquier</i>	1
CELL-MATRIX INTERACTIONS AND CELL ADHESION DURING DEVELOPMENT, <i>Peter Ekblom, Dietmar Vestweber, and Rolf Kemler</i>	27
SPATIAL PROGRAMMING OF GENE EXPRESSION IN EARLY <i>DROSOPHILA</i> EMBRYOGENESIS, <i>Matthew P. Scott and Patrick H. O'Farrell</i>	49
CELL ADHESION MOLECULES IN THE REGULATION OF ANIMAL FORM AND TISSUE PATTERN, <i>Gerald M. Edelman</i>	81
CORE PARTICLE, FIBER, AND TRANSCRIPTIONALLY ACTIVE CHROMATIN STRUCTURE, <i>D. S. Pederson, F. Thoma, and R. T. Simpson</i>	117
THE ROLE OF PROTEIN KINASE C IN TRANSMEMBRANE SIGNALLING, <i>Ushio Kikkawa and Yasutomi Nishizuka</i>	149
PROTON-TRANSLOCATING ATPASES, <i>Qais Al-Awqati</i>	179
REGION-SPECIFIC CELL ACTIVITIES IN AMPHIBIAN GASTRULATION, <i>John Gerhart and Ray Keller</i>	201
T-CELL ACTIVATION, <i>H. Robson MacDonald and Markus Nabholz</i>	231
ANCHORING AND BIOSYNTHESIS OF STALKED BRUSH BORDER MEMBRANE PROTEINS: Glycosidases and Peptidases of Enterocytes and Renal Tubuli, <i>Giorgio Semenza</i>	255
COTRANSLATIONAL AND POSTTRANSLATIONAL PROTEIN TRANSLOCATION IN PROKARYOTIC SYSTEMS, <i>Catherine Lee and Jon Beckwith</i>	315
THE DIRECTED MIGRATION OF EUKARYOTIC CELLS, <i>S. J. Singer and Abraham Kupfer</i>	337
PROTEIN IMPORT INTO THE CELL NUCLEUS, <i>Colin Dingwall and Ronald A. Laskey</i>	367
G PROTEINS: A FAMILY OF SIGNAL TRANSDUCERS, <i>Lubert Stryer and Henry R. Bourne</i>	391
	vii

viii	CONTENTS (<i>continued</i>)	
	MICROTUBULE-ASSOCIATED PROTEINS, <i>J. B. Olmsted</i>	421
	STRUCTURE AND FUNCTION OF NUCLEAR AND CYTOPLASMIC RIBONUCLEOPROTEIN PARTICLES, <i>Gideon Dreyfuss</i>	459
	MECHANISM OF PROTEIN TRANSLOCATION ACROSS THE ENDOPLASMIC RETICULUM MEMBRANE, <i>Peter Walter and Vishwanath R. Lingappa</i>	499
	REGULATION OF THE SYNTHESIS AND ASSEMBLY OF CILIARY AND FLAGELLAR PROTEINS DURING REGENERATION, <i>Paul A. Lefebvre and Joel L. Rosenbaum</i>	517
	INDEXES	
	Subject Index	547
	Cumulative Index of Contributing Authors, Volumes 1-2	557
	Cumulative Index of Chapter Titles, Volumes 1-2	558

Exhibit 41

Mutational Analysis of the Primary Substrate Specificity Pocket of Complement Factor B

ASP²²⁶ IS A MAJOR STRUCTURAL DETERMINANT FOR P₁-ARG BINDING*

(Received for publication, July 30, 1999, and in revised form, October 12, 1999)

Yuanyuan Xu^{‡§}, Antonella Circolo[‡], Hua Jing^{||}, Yue Wang[‡], Sthanam V. L. Narayana^{||},
and John E. Volanakis^{‡**}

From the [‡]Division of Clinical Immunology and Rheumatology, Department of Medicine and ^{||}Center for Macromolecular Crystallography, University of Alabama at Birmingham, Birmingham, Alabama 35294, the ^{||}Center for Blood Research, Harvard Medical School, Boston, Massachusetts 02138, and the ^{**}Biomedical Sciences Research Center "A. Fleming," Vari 166 72, Greece

Factor B is a serine protease, which despite its trypsin-like specificity has Asn instead of the typical Asp at the bottom of the S₁ pocket (position 189, chymotrypsinogen numbering). Asp residues are present at positions 187 and 226 and either one could conceivably provide the negative charge for binding the P₁-Arg of the substrate. Determination of the crystal structure of the factor B serine protease domain has revealed that the side chain of Asp²²⁶ is within the S₁ pocket, whereas Asp¹⁸⁷ is located outside the pocket. To investigate the possible role of these atypical structural features in substrate binding and catalysis, we constructed a panel of mutants of these residues. Replacement of Asp¹⁸⁷ caused moderate (50–60%) decrease in hemolytic activity, compared with wild type factor B, whereas replacement of Asn¹⁸⁹ resulted in more profound reductions (71–95%). Substitutions at these two positions did not significantly affect assembly of the alternative pathway C3 convertase. In contrast, elimination of the negative charge from Asp²²⁶ completely abrogated hemolytic activity and also affected formation of the C3 convertase. Kinetic analyses of the hydrolysis of a P₁-Arg containing thioester by selected mutants confirmed that residue Asp²²⁶ is a primary structural determinant for P₁-Arg binding and catalysis.

Complement is a major effector system of host defense. Activation of complement leads to the generation of protein fragments and protein-protein complexes that mediate acute inflammatory responses, phagocytosis and killing of pathogens, and regulation of adaptive immune responses. Activation-associated production of biologically active protein fragments is catalyzed by a group of eight atypical complement serine proteases (SPs)¹ of the chymotrypsin superfamily (1). Understand-

ing the structural basis for the highly restricted proteolytic activity of these SPs is an important first step toward pharmacologic control of complement activation (2).

Members of the chymotrypsin family have very similar three-dimensional structures but distinct substrate specificities. To a great extent specificity is determined by the side chains of the amino acid residues that line up the primary substrate specificity pocket (S₁ site). The pocket has three walls formed by residues 189–195, 214–220, and 225–228 (chymotrypsinogen numbering has been used for all SPs or SP domains throughout this paper) (3). The presence at the bottom of the pocket of Asp¹⁸⁹ endows trypsin with preference for positively charged Arg and Lys residues (4, 5), whereas in chymotrypsin the specificity for bulky aromatics is largely determined by Ser¹⁸⁹ (6). Residues at position 216 and 226 also contribute to substrate specificity (7). All complement SPs exhibit trypsin-like specificity for positively charged Arg residues and all have an Asp at position 189, except for factor B and C2 (Fig. 1).

Factor B and C2 are structurally similar modular proteins that play a central role in complement activation by providing the catalytic subunits of two key enzymes, namely the C3/C5 convertases of the alternative and the classical pathway, respectively. Complement convertases cleave the same single peptide bonds in C3 and C5. In addition to having Asn and Ser, respectively, instead of Asp at position 189, factor B and C2 also lack the highly conserved free N-terminal sequence of SPs. In typical SPs, the N-terminal sequence constitutes an essential structural element largely responsible for the transition from zymogen to active enzyme (8). Full expression of the proteolytic activities of factor B and C2 only occurs in the context of the complexes, C3bBb(C3b) and C4b2a(C3b), respectively (9). The SP domain resides in the C-terminal half of Bb or C2a and is preceded by a von Willebrand factor type A module (VWFA) which is noncovalently associated with C3b or C4b, respectively, in a Mg²⁺-dependent manner. These atypical structural features of factor B and C2 indicate a novel activation mechanism and probably also a distinct substrate binding arrangement at the primary specificity pocket.

In addition to their natural protein substrates C3 and C5, factor B and C2 and their fragments Bb and C2a hydrolyze a small number of C3- and C5-like synthetic substrates (11–14). Overall, C3-like substrates are considerably more reactive than C5-like substrates. However, even toward their best substrates, the k_{cat}/K_m values of factor B, Bb, C2, and C2a are

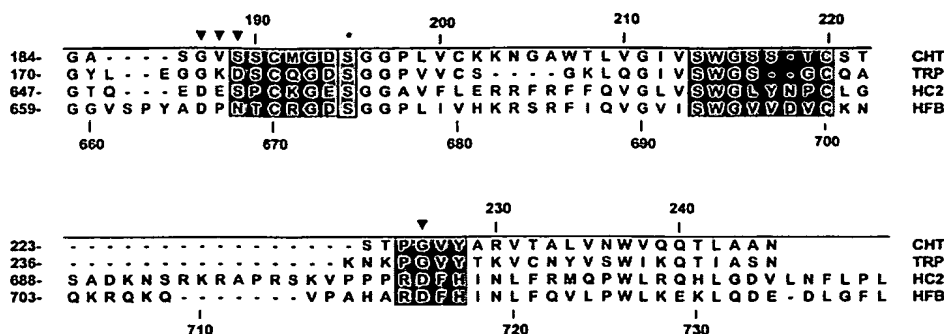
* This work was supported by National Institutes of Health Grants AI21067 (to J. E. V.), NIAMS, National Institutes of Health Grant P60 AR20614 R-3 (to Y. X.), and National Institutes of Health Grant AI39818 (to S. L. V. N.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: THT, Rm. 437, 1900 University Blvd., Div. of Clinical Immunology and Rheumatology, Dept. of Medicine, Birmingham, AL 35294. Tel.: 205-975-6241; Fax: 205-934-2126; E-mail: rheu019@uabdpd.dpo.uab.edu.

¹ The abbreviations used are: SP, serine protease; B-SP, the factor B serine protease domain; cCOLL, fiddler crab collagenase; CCP, complement control protein module; CHO, Chinese hamster ovary; CoVF, cobra venom factor; EC3b, erythrocytes sensitized with C3b; hNELA, human neutrophil elastase; hPRO3, human protease 3; mAb, mono-

clonal antibody; SBzl, thiobenzyl; VWFA, von Willebrand factor type A module; wt, wild type; Z, benzyloxycarbonyl; PAGE, polyacrylamide gel electrophoresis.

FIG. 1. Alignment of partial amino acid sequences of factor B, C2, chymotrypsin, and trypsin. Residues that form the walls of the primary specificity pocket are shaded. The catalytic triad residue Ser¹⁹⁵ is boxed and marked by an asterisk. Arrows indicate residues targeted for site-directed mutagenesis. Numbers at the top are for residues of the chymotrypsinogen sequence and those at the bottom are for the factor B sequence. CHT, bovine chymotrypsin; TRP, bovine trypsin; HC2, human C2; HFB, human factor B.



about 3 orders of magnitude lower than the $7.8 \times 10^6 \text{ s}^{-1} \text{ M}^{-1}$ value measured under the same conditions for the hydrolysis of the most reactive thioester by trypsin (14). By comparison, the catalytic efficiency (k_{cat}/K_m) of C3bBb for C3 cleavage was reported to be $3.1 \times 10^5 \text{ s}^{-1} \text{ M}^{-1}$ (10). No natural serine protease inhibitor has been found for factor B or C2 and regulation of the proteolytic activity of C3 convertases is effected largely through control of the assembly and decay of the bimolecular complexes. The structural correlates of the low esterolytic activity and extremely restricted substrate specificity as well as the conformational change(s) associated with zymogen activation are not understood. Determination of the structure of the factor B serine protease domain (B-SP) at 2.1-Å resolution has revealed the expected chymotrypsin fold but also unique features of surface loops and of the oxyanion hole.² The backbone conformation of the S₁ pocket is similar to that of trypsin, but there are substitutions of functionally important residues. In this study we used site-directed mutagenesis to analyze possible effects of the factor B-specific residues on the assembly and activity of the C3 convertase. The data indicate that Asp²²⁶ is a primary structural determinant of P₁-Arg binding and that the native conformation of Asp²²⁶ and Asn¹⁸⁹ are important determinants for C3 cleavage.

EXPERIMENTAL PROCEDURES

Construction of Mutant Factor B cDNA—The factor B cDNA clone BHL4-1 (15) in the expression vectors pRC/CMV or pCDNA3 (Invitrogen, Carlsbad, CA) was used as wild type (wt) template in site-directed mutagenesis. Factor B mutant cDNA constructs were obtained by the method of Zollar and Smith (16) as modified by Kunkel (17). Alternatively, the QuikChange Site-directed mutagenesis kit (Stratagene, La Jolla, CA) was used according to the manufacturer's protocol. All cDNA constructs of mutant factor B were verified by restriction mapping and dideoxynucleotide sequencing (18) of the region around the mutation. Oligonucleotides were synthesized by the phosphoramidite method (19), using a DNA/RNA synthesizer (Model 394 Applied Biosystems, Foster City, CA).

Expression of wt and Mutant Factor B cDNA—Transient transfection of COS cells with 30–40 µg of cDNA was performed by electroporation as described (20). Cell culture supernatant containing secreted factor B proteins was harvested 72–90 h after transfection. Cell debris was removed by centrifugation and the supernatant was stored frozen at –80 °C in small aliquots. The concentration of recombinant factor B in the medium was measured by enzyme-linked immunosorbent assay (15), using a rabbit anti-human Bb IgG (50 µg/ml) as capturing antibody and the mouse anti-Ba monoclonal antibody (mAb) HA4-ID5 (1.5 µg/ml) as reporter. The assay was developed with 1:1000 dilution of affinity-purified goat anti-mouse IgG1 alkaline phosphatase conjugate (Southern Biotechnology Associates, Birmingham, AL) and Sigma substrate 104 (Sigma). Color development was measured at 405 nm. The concentration of factor B was calculated from a standard curve constructed using human serum of known factor B concentration. The sensitivity of the assay was approximately 1–2 ng/ml and the concentration of specific protein in the culture medium ranged from 0.3 to 2 µg/ml.

To obtain large amounts of recombinant proteins, stable transfection of Chinese hamster ovary cells (CHO-K1, ATCC) was carried out with selected mutants by a modification of a previously described method (21). CHO-K1 cells were maintained in Ham's F-12 (Cellgro, Herndon, VA) supplemented with 10% heat-inactivated fetal bovine serum (Life Technologies, Grand Island, NY), and 2 mM glutamine at 37 °C in a humidified, 5% CO₂ incubator. Forty micrograms of each CsCl-purified plasmid DNA was transfected into 4–6 × 10⁶ CHO-K1 cells by electroporation as described (21). Selection of neomycin-resistant cells was started 72 h after transfection with 750 µg of G418 (Cellgro) per ml of the above medium. Subcloning of the G418-resistant cells was performed approximately 7 days after initiating selection by limiting dilution of cells at 0.8 cell/well in 96-well tissue culture plates. Clones were allowed to grow in G418-containing medium with 15% heat-inactivated fetal bovine serum for 10–12 days before screening for factor B production by enzyme-linked immunosorbent assay. The highest producing wt and mutant factor B clones were selected, expanded, and adapted to large-scale production by growing in suspension culture for 2 weeks. Protein purification was facilitated by culturing cells in ExCell 301 serum-free medium (JRH Bioscience, Lenexa, KS) supplemented with 0.5–2% fetal bovine serum, 2 mM glutamine, and 200 µg/ml G418.

Purification of Recombinant wt and Mutant Factor B—One to two liters of the stably transfected CHO cell culture medium were harvested, concentrated to approximately 150 ml, and applied to a 30-ml column of CM Sephadex C-50 equilibrated with 0.1 M sodium acetate, 20 mM ε-amino-*n*-caproic acid, 20 mM EDTA, pH 6.5. Factor B was eluted with a gradient of 0–0.2 M NaCl in the starting buffer. For further purification, factor B-containing pools were dialyzed against 20 mM Tris-HCl, pH 8.0, and subjected to fast protein liquid chromatography, using a Mono-Q column (Amersham Pharmacia Biotech). Factor B was eluted with a gradient of 0–0.3 M NaCl in the starting buffer. For some mutants Mono-Q chromatography was repeated. Purity of factor B proteins assessed by 10% SDS-PAGE was between 80 and 95%.

Reactivity of Factor B Mutants with Module-specific MAb—Two anti-Ba mAbs, HA4–1D5 (a subclone of HA4–1A) and FD3–20, and an anti-Bb mAb, HA4–15, were described previously (22). The mAb 6B3.3 was raised by using as antigen recombinant factor B VWFA module expressed in *Escherichia coli*. Reactivity of factor B mutants with these mAbs was examined by enzyme-linked immunosorbent assay similar to that described above. The same rabbit anti-human Bb IgG antibody was used in the solid phase, and each of the four mAbs was used as detectant at a concentration of 1.5 µg/ml. The assay was developed with goat anti-mouse IgG + IgM alkaline phosphatase conjugate (Jackson ImmunoResearch Laboratory, Inc., West Grove, PA) and phosphatase substrate Sigma 104. Values obtained for each mAb were normalized to those measured for HA4–1D5 and represent the average of two separate experiments.

Solid-phase Cobra Venom Factor (CoVF) Binding Assay—Binding of wt and mutant factor B to CoVF was determined by enzyme-linked immunosorbent assay as described (23). Culture medium from transfected COS cells containing wt or mutant factor B was dialyzed against half-strength veronal-buffered saline (0.5 × veronal-buffered saline, 2.5 mM sodium 5, 5-diethylbarbiturate, pH 7.4) containing 5 mM MgCl₂ at 4 °C overnight. Serial dilutions of factor B in the same buffer were then added to microplates coated with CoVF (Quidel, San Diego, CA). Binding of factor B to CoVF was allowed to occur in the absence or presence of 1.5 µg/ml factor D at 37 °C for 2 h. Bound factor B or Bb were detected with rabbit anti-Bb IgG (50 µg/ml) and goat anti-rabbit IgG alkaline phosphatase conjugate. Results represent the average values of two separate experiments.

CoVF-mediated Factor B Cleavage by Factor D—COS cells (4–6 ×

² Jing, H., Xu, Y., Carson, M., Moore, D., Macon, K. J., Delucas, L. J., Volanakis, J. E., and Narayana, S. V. L. (2000) *EMBO J.* 20, in press.

10⁶) were transiently transfected by electroporation with wt or mutant factor B cDNA as described above. The cells were metabolically labeled 72 h later in 1 ml of Dulbecco's modified Eagle's medium without methionine, supplemented with 250 μ Ci of [³⁵S]Met (specific activity ~ 1000 Ci/mmol, Amersham Pharmacia Biotech or ICN Radiochemical, Irvine, CA.) for 30 min and chased with cold methionine in Dulbecco's modified Eagle's medium supplemented with 10% heat-inactivated fetal bovine serum. After a 3-h chase, 650- μ l aliquots of the culture supernatants were collected, supplemented with 25 mM Tris-HCl, pH 7.4, 2.5 mM MgCl₂, and incubated for 2 h at 37 °C with factor D (300 and 2 ng) in the absence or presence of 5 μ g of CoVF. Labeled factor B and Bb were immunoprecipitated by using rabbit anti-Bb IgG antibody and *Staphylococcus aureus* protein A and analyzed by SDS-PAGE as described (24). To assess factor B cleavage, gel slices corresponding to the autoradiographed bands and blank spaces were cut and digested with 15% H₂O₂ at 56 °C overnight. The blank gel cuts were used to subtract background radioactivity. The released radioactivity was measured with Bio Safe II scintillation fluid (RPI, Mount Prospect, IL) in an LKB liquid scintillation counter (Model 1215 LKB, Gaithersburg, MD) (25).

Factor B Hemolytic Assay—Sheep blood erythrocytes carrying C3b (EC3b) were prepared as described (22), by using freshly purified human factor B (22), factor D (26), and C3 (27). Serial dilutions of culture medium containing wt or mutant factor B were added to 7.5 \times 10⁶ EC3b, 12.5 ng of factor D, and 125 ng of properdin (Sigma) in a total volume of 150 μ l in 0.5 \times veronal-buffered saline containing 2.5% dextrose, 2.5 mM MgCl₂, 10 mM EGTA, and 0.1% gelatin. Formation of C3 convertase, C3bBb(P), was carried out at 30 °C for 30 min. Then, 0.5 ml of guinea pig serum diluted 1:40 with 10 mM EDTA in veronal-buffered saline was added as source of C3 to C9 and the reaction mixture was incubated for 1 h at 37 °C. Percent lysis and hemolytic units/ μ g were calculated as described (28). Values of specific hemolytic activity of each mutant were normalized to that of wt factor B and represent the mean \pm S.E. of at least three independent determinations, each performed in duplicate.

C3 Cleavage Assay—C3 was freshly isolated from plasma of a normal individual as described (27) except that a final chromatographic step using hydroxyapatite fast protein liquid chromatography (Amersham Pharmacia Biotech) was added. Purified wt or mutant factor B (50 ng) was mixed with C3 (75 ng) with or without 150 ng of CoVF and 12.5 ng of factor D in a total volume of 25 μ l of 25 mM Tris-HCl, pH 7.4, containing 75 mM NaCl and 5 mM MgCl₂. After incubating at 37 °C for 1 h, 10 μ l of each reaction mixture was analyzed on 7.5% SDS-PAGE. C3 and C3 fragments were detected on Western blots by using goat anti-human C3 IgG (Cappel, Durham, NC) and affinity-purified rabbit anti-goat IgG F(ab')₂ horseradish peroxidase conjugate (ICN). The ECL luminescent detection system (Amersham Pharmacia Biotech) was utilized to visualize C3 polypeptide chains following the manufacturer's protocol. The amount of C3 conversion was determined by scanning α and α' chain using ScanMaker 5 scanner (MicroTek Lab, Inc., Redondo Beach, CA) and band intensity was quantified using software NIHImage1.58.

Esterolytic Assays—The rate of hydrolysis of Z-Lys-Arg-SBzl (Peninsula Laboratories Belmont, CA) was measured by a modification of the method of Kam *et al.* (14). Assays were carried out in microplate wells. The B-SP was expressed by Sf9 insect cells infected by recombinant baculovirus and isolated from the serum-free Excell 401 media using Bio-Rex 70 and Mono S ion exchange chromatography.² The recombinant B-SP consists of a vector-derived tripeptide Ala-Asp-Pro at the N terminus and the C-terminal 295 amino acid residues of factor B. Purified factor B or B-SP (0.11–0.2 μ M) was added to 0.08 to 0.8 mM Z-Lys-Arg-SBzl and 1.6 mM Ellman's reagent 5,5-dithiobis-(2-nitrobenzoic acid) (Sigma) in 250 μ l of 0.1 M HEPES, pH 7.5, containing 0.5 M NaCl and 16% Me₂SO. Factor B was omitted from control wells used for measuring background hydrolysis of the substrate. Esterolytic rates were measured kinetically for 15 min by using a V_{max} kinetic microplate reader (Molecular Devices, Menlo Park, CA). Kinetic constants were determined by the Lineweaver-Burk method based on at least five substrate concentrations. Correlation coefficients in all cases were greater than 0.98.

RESULTS

To understand the structural implications of the unique factor B residues in and around the primary specificity pocket, the serine protease domain (B-SP) was expressed using a baculovirus system and its crystal structure determined at 2.1-Å resolution by multiple isomorphous and molecular replacement methods.² As expected, B-SP was found to display a chymo-

trypsin-like, two β -barrel structural fold. In the active center, the catalytic triad residues, Asp¹⁰², His⁵⁷, and Ser¹⁹⁵, and the nonspecific substrate-binding site (Ser-Trp-Gly^{214–216}) have typical serine protease configurations (Fig. 2). However, the oxyanion hole displays a zymogen-like conformation due to the inward orientation of the carbonyl oxygen atom of Arg¹⁹², the backbone of which together with those of Cys¹⁹¹, Gly¹⁹³, and Asp¹⁹⁴ form a single-turn 3₁₀ helix. The three walls of the primary specificity pocket are formed by residues 189–195, 214–220, and 225–228. The backbones of these residues, except for the single-turn helix, can be superposed on those of the corresponding residues of trypsin. Asn¹⁸⁹ is located at the bottom of the pocket, replacing the highly conserved Asp of other SPs with trypsin-like substrate specificity. However, the side chain of Asp²²⁶, which replaces Gly²²⁶ of trypsin, extends toward the bottom of the pocket which suggests that it may be directly involved in binding the P₁-Arg of the substrate substituting for Asp¹⁸⁹ of other trypsin-like SPs. An Asp residue also replaces a conserved Gly of other SPs at position 187. Asp¹⁸⁷ of factor B is located directly beneath the pocket and forms a salt bridge with Lys¹⁶³. To investigate the possible participation of the three residues, Asp¹⁸⁷, Asn¹⁸⁹, and Asp²²⁶, in substrate binding and catalysis, factor B mutants at these positions were constructed and assayed. In addition, the functional role of Pro¹⁸⁸, not found at this position in other SPs, was also assessed. In most cases, two independent clones for each mutant were expressed and analyzed to avoid artifactual results. In all cases, results of functional analysis of the two clones of each mutant were consistent. This suggested that functional differences from the wt resulted from the amino acid substitution at the mutation sites.

Reactivity of Factor B Mutants with Module-specific MAbs—To probe for possible effects of the mutations on the overall structure of the molecule, we tested the reactivity of the mutants with a panel of module-specific mAbs. The anti-Bb mAb HA4–15 (22) has been shown to recognize an epitope on the SP domain (data not shown). MAbs FD3–20 (anti-CCP1–3) and HA4–1D5 (anti-CCP2) bind to distinct epitopes on the Ba fragment (29), while 6B3.3 (γ 1, κ) recognizes an epitope on the VWFA module at or near the C3b-binding site (data not shown). We did not observe substantial differences in the reactivity of the mutants with the four mAbs (data not shown), suggesting that all epitopes tested are retained in their native conformation.

Formation of the CoVFB and CoVFBb Complexes—Expression of proteolytic activity by the factor B SP domain requires binding of factor B to C3b and its proteolytic cleavage by factor D. Introducing mutations in the SP domain could alter C3b binding and/or susceptibility to factor D cleavage, although these functions have been assigned to distal parts of the molecule, namely, the CCP and the VWFA modules (1). We examined the ability of factor B mutants to form the CoVFB and CoVFBb complexes. Choice of CoVF over C3b was dictated by the much longer half-life of the complexes, which facilitates detection. All mutants showed dose-dependent binding to CoVF in the absence (data not shown) and presence (Fig. 3) of factor D. Enhancement of binding to CoVF was observed in the presence of factor D for all mutants. Factor B carrying single mutations at positions 187 or 189 had essentially the same binding activity as wt factor B, except for the D187Y mutant, which only formed about half as much CoVFBb as wt factor B. In the D226 panel of mutants, surprisingly only D226N had wt binding activity. The same substitution combined with N189D resulted in 50% reduction of binding to CoVF compared with either the D226N or N189D mutant. The trypsin-like mutation D226G alone or in combination with the N189D mutation

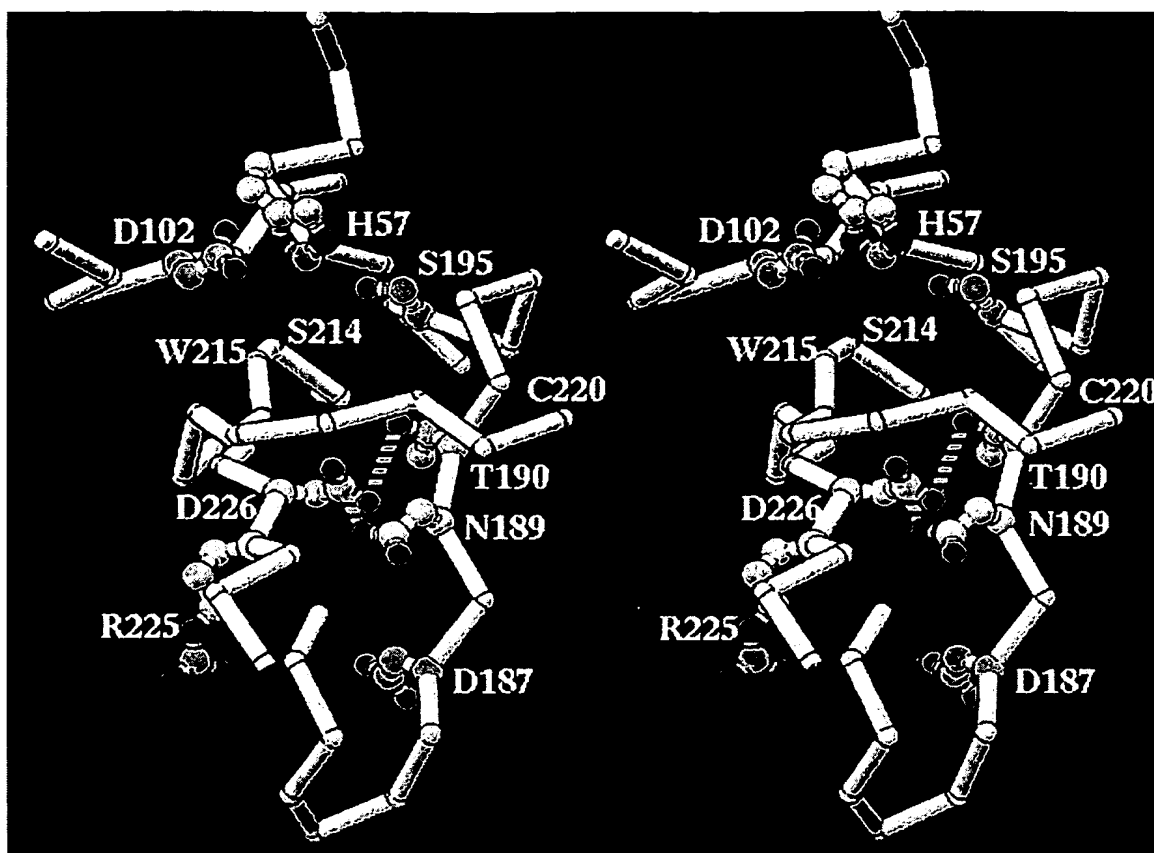


FIG. 2. Stereoview of the active center of the factor B serine protease domain. The side chains of the catalytic triad residues and of selected residues lining the S_1 pocket are shown. Hydrogen bonds between the carboxyls of Asp²²⁶ and the side chains of Asn¹⁸⁹ and Thr¹⁹⁰ are shown by dashed lines.

caused 60 and 87% reduction, respectively, in CoVFBb complex formation. Similar reductions in CoVF binding ability of the mutants was also observed without factor D cleavage (data not shown). The results suggested that, with the exception of the D226N mutation, substitutions at position 226 affect initial binding of factor B to CoVF thus sensitivity to factor D proteolysis, since binding is a prerequisite for factor B cleavage. In a more direct factor B cleavage assay, conversion of biosynthetically labeled factor B to Bb by factor D in the presence of CoVF was analyzed by SDS-PAGE and autoradiography (Fig. 4). The results correlated well with the binding data. Mutant D226N was as sensitive to factor D cleavage as wt factor B. Mutants D226N/N189D, D226G, and D226G/N189D were less susceptible to factor D with conversion to Bb estimated at 53, 27, and 16%, respectively, of that of wt factor B at the high concentration of factor D. The combined results suggest that although the overall structural integrity of the mutants was preserved, as indicated by equivalent reactivity with the module-specific mAbs, amino acid substitutions in the SP domain apparently affected CoVF/C3b binding, which is mediated by sites on the other two domains of the molecule.

Hemolytic Activity of Factor B Mutants—The effects of the mutations on the ability of factor B to cleave/activate C3 and C5 were assessed by a hemolytic assay. The hemolytic activity of the mutants relative to that of wt factor B is illustrated in Fig. 5. Elimination of the negative charge of Asp¹⁸⁷ in mutants D187A, D187N, and D187S resulted in 50–60% loss of hemolytic activity. Substitution of Tyr at the same position caused a more pronounced decrease in hemolytic activity, approximately 80%. The data suggest that the bulky hydrophobic side chain of

Tyr is not favored and that full expression of factor B hemolytic activity requires the salt-bridging conformation of Asp¹⁸⁷. Ala mutation at position 188 in the mutant P188A did not have significant effect on the hemolytic activity.

As revealed in the crystal structure, Asn¹⁸⁹ and the side chain of Asp²²⁶ are located at the bottom of the primary specificity pocket and appear to be accessible to the P₁-Arg of the substrate (Fig. 2). Replacement of Asn¹⁸⁹ with charged residues, either Asp or Lys, reduced hemolytic activity by 95%, while the Ala mutant retained approximately 30% of wt activity. Although eliminating the negative charge from Asp²²⁶ in the D226N mutant did not affect the assembly of the CoVFBb complex (Fig. 3), it completely abrogated the C3/C5 convertase activity. Replacement of the same residue with Gly present in trypsin also resulted in complete loss of hemolytic activity. Again the loss of hemolytic activity was out of proportion to the only moderately reduced ability to form the CoVFBb complex (Fig. 3). Attempts to construct a trypsin-like pocket by reassigning the negative charge to position 189 in the double mutants D226N/N189D and D226G/N189D failed to restore factor B hemolytic activity, despite the residual CoVF binding activity (Figs. 3 and 5). The hemolytic data strongly indicate that Asp²²⁶ plays a critical and highly specialized role in the expression of C3/C5 convertase activity by factor B. Residue Asn¹⁸⁹ and Asp¹⁸⁷ are also of importance for expression of factor B-dependent proteolytic activity. In contrast, the Pro residue at position 188 has no apparent functional role and likely serves as spacer between structurally crucial residues.

C3 Cleavage Assay—Decrease of the factor B hemolytic activity could reflect a defect of C3 and/or C5 cleavage. The effects

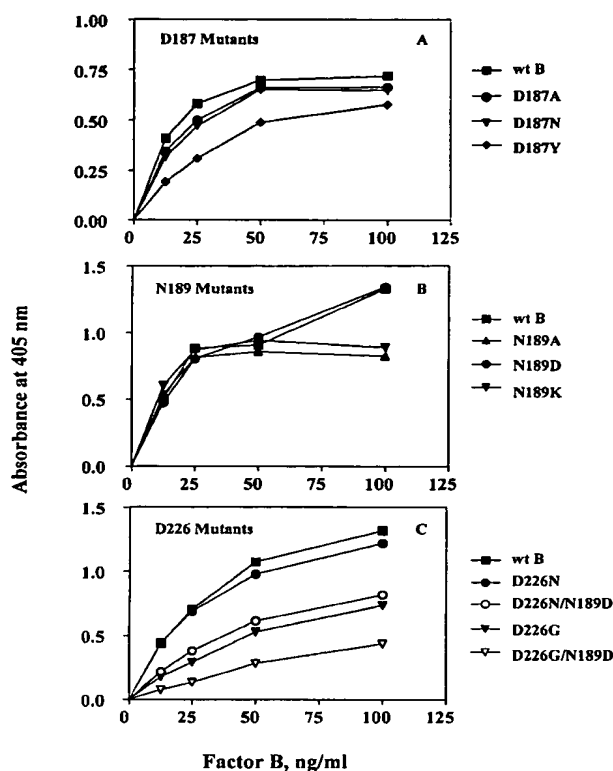


FIG. 3. Assembly of solid-phase CoVFBb complex by wt and mutant factor B. Microtiter plates were coated with CoVF (10 $\mu\text{g}/\text{ml}$). Serial dilutions of wt and mutant factor B in culture supernatants of transfected COS cells were added and incubated with factor D (1.5 $\mu\text{g}/\text{ml}$) at 37 $^{\circ}\text{C}$ for 2 h. CoVF-bound Bb fragments were detected by using rabbit anti-human Bb IgG and goat anti-rabbit IgG as detailed under "Experimental Procedures." Symbols are: A, \blacksquare , wt B; \bullet , D187A; \blacktriangledown , D187N; \blacklozenge , D187Y; B, \blacksquare , wt B; \blacktriangle , N189A; \bullet , N189D; \blacktriangledown , N189K; C, \blacksquare , wt B; \bullet , D226N; \circ , D226N/N189D; \blacktriangledown , D226G; ∇ , D226G/N189D.

of the mutations on C3 proteolytic activity were assessed by a direct cleavage assay. Wt factor B and selected mutants were permanently expressed in CHO cells and purified. Fluid-phase C3 convertases were formed with CoVF in the presence of factor D. Conversion of C3 to C3a and C3b was assessed by the appearance of the α' chain of C3b on SDS-PAGE (Fig. 6). As shown, under the experimental conditions used, wt factor B converted 45% of α to α' chain, while there was no conversion observed in controls not containing CoVF and factor D. The N189A mutant demonstrated 37% of wt proteolytic activity. This is consistent with the expression of 29% of wt hemolytic activity by this mutant (Fig. 5). As expected from the lack of hemolytic activity, there was no detectable C3 cleavage by the D226N and D226N/N189D mutants even after prolonged exposure of the film. However, there was trace amount of α chain cleavage by the N189D mutant, seen more clearly after long exposure of the film. The C3 cleavage study demonstrated that at least for the factor B mutants tested loss of hemolytic activity could be attributed to loss of proteolytic activity for C3.

Esterolytic Activity—Because C3 is a large protein substrate, extensive molecular contacts with C3b-bound Bb are probably required for its proteolysis. Hydrolysis of small synthetic thioester substrates containing Arg at the P_1 site could provide further insights into substrate recognition. In the present study we chose Z-Lys-Arg-SBzl as substrate because it was shown to be the most reactive among the P_1 Arg-containing C3 or C5-like substrates tested by Kam *et al.* (14). The catalytic efficiency (k_{cat}/K_m) of recombinant wt factor B was 1135 M^{-1}

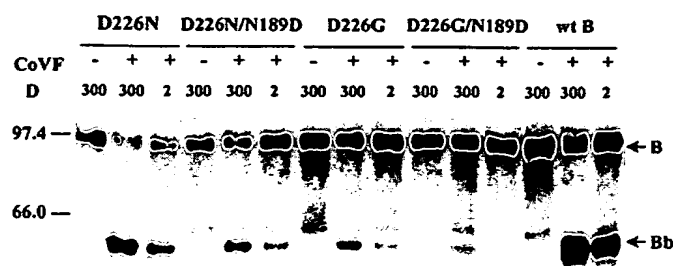


FIG. 4. Cleavage of CoVF-bound factor B by factor D. [^{35}S]Met-labeled wt and Asp 226 factor B mutants secreted by transiently transfected COS cells were incubated with two different concentrations of factor D in the presence of 5 μg of CoVF for 2 h at 37 $^{\circ}\text{C}$ or with the high concentration of factor D in the absence of CoVF as control. After incubation, immunoprecipitation was performed by using a rabbit anti-human Bb IgG and *S. aureus* protein A. Immunoprecipitates were washed and subjected to 7.5% SDS-PAGE and autoradiography. Positions and molecular mass of marker proteins are given on the left.

s^{-1} (Fig. 7) which is similar to the 1370 $\text{M}^{-1} \text{s}^{-1}$ value reported previously for native factor B (14). The recombinant B-SP had k_{cat}/K_m of 198 $\text{M}^{-1} \text{s}^{-1}$, which is 5.7 times lower than that of intact factor B. Measurement of individual kinetic parameters showed that the decreased k_{cat}/K_m of B-SP was mainly due to a 4-fold increase in K_m . Of the mutants tested, D226N showed 50-fold slower catalytic rate than wt factor B. However, placement of a negative charge at position 189 on the D226N background partially restored esterolytic activity. As shown, the k_{cat}/K_m of the double mutant D226N/N189D was about 10-fold higher than that of D226N. As indicated by the lower than wt factor B k_{cat} and unaltered K_m , decreased catalytic efficiency of these two mutants could be directly attributed to the decreased catalytic rate. These results strongly suggest that the negatively charged Asp 226 determines binding specificity and catalytic efficiency for the substrate Z-Lys-Arg-SBzl. Substitutions of Asp or Ala for Asn 189 in N189D and N189A caused 2.7- and 6.6-fold lower activity, respectively. Although N189A factor B had slightly lower esterolytic activity than N189D factor B, it had substantially higher proteolytic activity for C3 (Fig. 6). Our findings demonstrated that in addition to Asp 226 , Asn 189 also participates in substrate recognition and in determining specificity for C3. Apparently, the structural configuration of residues Asp 226 and Asn 189 of factor B is critical for recognition and cleavage of C3 and C5.

DISCUSSION

Determination of the structure of the SP domain of factor B revealed a number of novel insertions and deletions compared with typical SPs and also certain unique structural features of the catalytic apparatus, especially in the primary specificity pocket (data not shown). In the present study, mutational analysis of factor B residues in and around the primary specificity pocket was performed to investigate structural correlates of substrate recognition at the S_1 site. The results are discussed in light of the large amount of available information on SP specificity.

Our results clearly demonstrate that Asp 226 of factor B is a critical structural determinant for substrate binding and catalysis, substituting for Asp 189 of other SPs with trypsin-like specificity. Functional analysis of the D226N mutant provided the most clear-cut results. The observed loss of esterolytic and proteolytic activity of this mutant could be attributed solely to a catalytic defect resulting from inappropriate engagement of the P_1 -Arg in the S_1 site, while other functional sites necessary for the proteolytic activation and substrate binding appeared to be well preserved. A sharp 50-fold decrease in catalytic rate (k_{cat}) indicates that a negative charge at the bottom of the

FIG. 5. Hemolytic activity of factor B mutants. EC3b (1.5×10^7) were incubated with serial dilutions of wt and mutant factor B in culture medium of transfected COS cells, factor D (12.5 ng), and properdin (125 ng) at 30 °C for 30 min. Hemolysis was allowed to occur at 37 °C for 1 h after addition of 1:40 dilution of guinea pig serum in EDTA buffer. For each mutant specific hemolytic activity (units/ μ g) was calculated and normalized to that of wt B. Each bar represents the average \pm S.E. of the results of at least three separate experiments performed in duplicate.

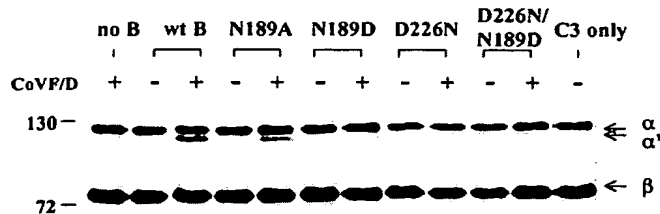
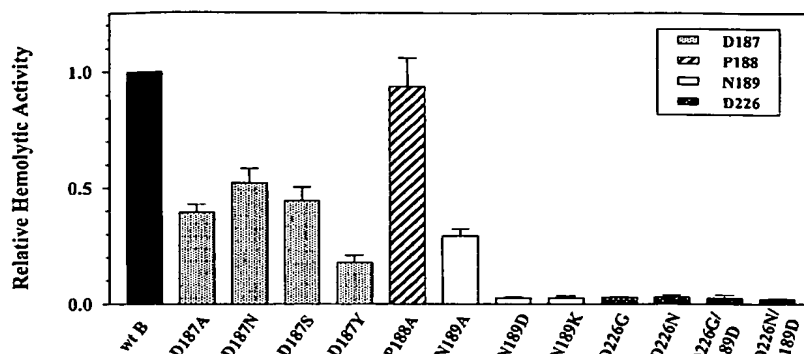


FIG. 6. Proteolytic activity of C3 convertases formed by CoVF and wt or mutant factor B. Wt or mutant factor B (50 ng) and C3 (75 ng) were incubated for 1 h at 37 °C with (+) or without (−) CoVF (150 ng) and D (12.5 ng). Aliquots of the reaction mixture were analyzed on 7.5% SDS-PAGE under reducing conditions. C3 polypeptide chains were detected on Western blots by using a goat anti-human C3 IgG. Positions and molecular mass of marker proteins are shown on the left. Positions of α , α' , and β chains of C3 are given on the right.

primary pocket is essential for efficient catalysis, but not for overall substrate binding affinity, because the K_m is not altered by the Asn substitution (Fig. 7). Apparently, hydrogen bond formation of the P_1 - P_3 residues to the nonspecific substrate-binding site, Ser-Try-Gly^{214–216}, and hydrophobic anchoring of the P_2 and P_3 side chains to S_2 and S_3 pockets, respectively, provide sufficient binding force. Also it seems likely that Asn²²⁶ provides additional binding energy, probably by hydrogen bonding with P_1 -Arg. However, positioning of the scissile bond relative to Ser¹⁹⁵ and the oxyanion hole through the putative hydrogen bonds may differ from that effected by the direct ionic contact made by Asp²²⁶ in wt factor B. Replacing Asp²²⁶ with Asn affected equally esterolytic and C3 proteolytic activity, although D226N factor B could form a CoVFBb complex. In a recent report Hourcade *et al.* (30) also found that substitution of various residues (Asn, Ala, Ser, and Tyr) for Asp²²⁶ caused severe reduction in proteolytic activity despite normally assembled C3bBb complex. It is of special interest that the conservative substitution of Glu for Asp²²⁶ also abrogated C3 proteolytic activity. This observation suggests that accurate positioning of the carbonyl group of P_1 -Arg of C3 relative to the nucleophilic Ser¹⁹⁵ O- γ and oxyanion hole can only be achieved by the native residue Asp²²⁶. A corresponding trypsin mutant, D189E, displayed 2–3 orders of magnitude decrease in catalytic efficiency (k_{cat}/K_m), associated with a 40-fold shift in the preference from Arg to Lys substrates relative to wt trypsin (31). Apparently, the additional methylene group distancing the carboxylate of trypsin D189E from the peptide backbone within the narrow S_1 pocket impeded the proper positioning of the side chain of Arg, which is longer and larger than that of Lys. The loss of C3 catalytic activity by D226E factor B (30) can probably be attributed to a similar spatial effect.

Another structural characteristic of the S_1 pocket of factor B is a hydrogen bonding network formed by the carboxyl oxygens

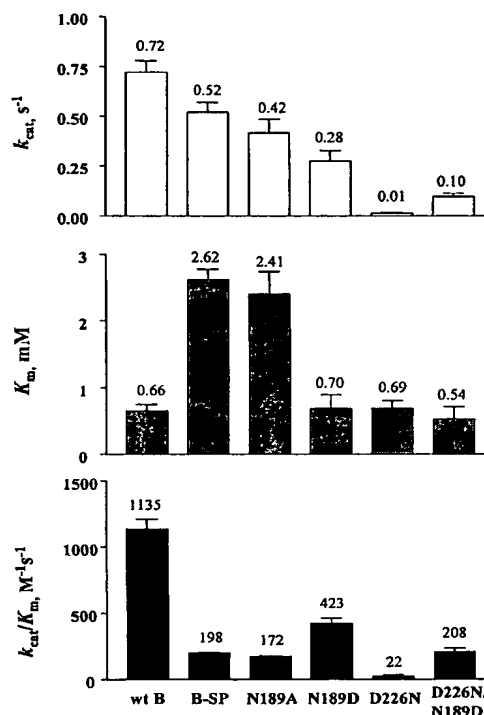


FIG. 7. Hydrolysis of synthetic thioester substrate by wt and mutant factor B and the factor B serine protease domain. Purified wt or mutant factor B or recombinant B-SP (113–200 nm) was incubated with Z-Lys-Arg-SBzl at concentration of 0.08–0.8 mM. Hydrolysis was measured at 25 °C in the presence of Ellman's reagent 5,5-dithio-bis-(2-nitrobenzoic acid) used as a chromogen of hydrolysis. Kinetic parameters were derived from Lineweaver-Burk plots. The values of individual parameters are the average \pm S.E. of at least three independent determinations.

of Asp²²⁶ and pocket residues Asn¹⁸⁹, Thr¹⁹⁰, and Arg²²⁵ (Fig. 2). This effectively reduces ionic bonding potential available for making contacts with P_1 -Arg of the substrate. On one hand, this distinct feature could possibly explain the overall low esterolytic activity of factor B, Bb (12–14), and B-SP (Fig. 7). On the other hand, it implies the need for additional bonding between P_1 -Arg and other pocket residues. The side chain of Asn¹⁸⁹ faces the carboxyl of Asp²²⁶ from the opposite wall and occupies a central position at the bottom of the specificity pocket. Although the position of the Asn¹⁸⁹ side chain is about 0.5–1.0 Å lower than that of Asp²²⁶, it appears accessible to the substrate. Our results indicate a supporting role for Asn¹⁸⁹ in substrate recognition and catalysis. Substitution of Ala, Asp, or Lys at this position caused substantial reduction or abrogation of hemolytic activity, which paralleled a similar reduction in C3

proteolytic activity (Figs. 5 and 6). The Ala substitution caused a decline in synthetic substrate binding affinity (K_m) and catalytic efficiency (k_{cat}/K_m), which strongly indicates participation of Asn¹⁸⁹ in substrate recognition. The amine group of the Asn¹⁸⁹ side chain may mediate P₁-Arg binding through a hydrogen bond. Absence of this potential binding force may compromise accurate register of P₁-Arg of C3 for catalysis. Substitution of a charged residue, Asp or Lys for Asn¹⁸⁹ in N189D and N189K, respectively, abrogates C3 proteolytic activity of the C3- or CoVF-bound Bb. Interestingly, the N189D mutant retains substantial esterolytic activity toward the synthetic substrate. These results suggest that the reconstructed S₁ pocket, with free carboxyls at positions 226 and 189, despite its altered geometry could register to the His⁵⁷-Ser¹⁹⁵ dyad, the Arg bond of the synthetic substrate but not that of C3. The free leading or leaving group of the synthetic substrate may account for the observed binding flexibility.

C2 and factor B have identical proteolytic specificity for single Arg peptide bonds of C3 and C5 so that their substrate-binding sites can be presumed to be very similar in geometry and chemical nature. Thus, it is not surprising that C2 has Asp and Ser at positions 226 and 189, respectively (Fig. 1). Besides factor B and C2, an acidic residue is also present at position 226 in a few additional members of the chymotrypsin family, namely fiddler crab collagenase (cCOLL) (32), human cathepsin G (CATG) (33), protease 3 (hPRO3) (34), and neutrophil elastase (hnELA) (35). In contrast to C2 and factor B these serine proteases display relatively broad substrate specificity. cCOLL and CATG recognize not only basic but also large hydrophobic side chains (32, 36). The Arg/Lys substrate preference is mainly attributed to the presence of Asp²²⁶/Gly¹⁸⁹ in cCOLL and of Glu²²⁶/Ala¹⁸⁹ in CATG within the S₁ pocket. The large and flexible S₁ pocket in cCOLL allows this enzyme to adjust to different shapes of the P₁ side chain. Removal of the negative charge from the cCOLL S₁ pocket in the D226G mutant resulted in a significant decrease of catalytic efficiency toward Arg/Lys substrates (37). Similarly to Asp²²⁶ in factor B and cCOLL, the corresponding Glu²²⁶ in human CATG has only one carboxyl oxygen available for substrate binding (33). This may be responsible for the relatively slow catalysis of substrates with P₁-Lys or Arg. However, the presence of a negatively charged residue at position 226 is not a sufficient condition for specificity for basic residues. Neither hPRO3 nor hnELA, both of which have an Asp²²⁶, recognizes a Lys or Arg-P₁ residue. The two enzymes display close similarity of their S₁ sites and cleave after small mostly hydrophobic residues, such as Leu/Ile (hnELA), Ala/Ser (hPRO3), and Val/Met (hnELA and hPRO3) (38). The presence of Ile and Val at position 190 of hPRO3 and hnELA, respectively, seems partially responsible for their substrate specificities. In hnELA, loss of specificity for basic residues has been attributed to inaccessibility of Asp²²⁶ that is shielded by Val¹⁹⁰ and Val²¹⁶. Similarly, Asp²²⁶ of hPRO3 is also shielded by Ile¹⁹⁰ and Val²¹⁶. Taken together, the data indicate that Arg/Lys substrate specificity is structurally determined not only by the presence but also by the accessibility of an acidic side chain at the base of the specificity pocket, positioned either at 189 or 226. The carboxyl oxygens of Asp²²⁶ or Glu²²⁶ seem less available to substrate than those of Asp¹⁸⁹ because of participation in hydrogen-bonding networks with residues on the wall of the pocket. This appears to be a distinct feature observed in factor B, the neutrophil elastases, and cCOLL.

Structural and functional consequences of altering the Asp¹⁸⁹ of trypsin have been examined by site-directed mutagenesis, kinetic, and crystallographic analysis (39). The negative charge was relocated to the opposite wall of the binding

pocket in rat trypsin mutant D189G/G226D. Kinetic analysis showed that, compared with wt trypsin, this relocation of the negative charge caused 10⁴- and 4.5 × 10²-fold decrease in catalytic efficiency (k_{cat}/K_m) toward P₁-Arg and -Lys containing substrates, respectively. The decrease resulted from a much sharper decline in k_{cat} for the Arg than the Lys substrates, whereas the binding affinity (K_m) for both substrates was equally reduced. The crystal structure of D189G/G226D trypsin in complex with inhibitors showed that in its new position, Asp interacts extensively with other residues in the pocket through hydrogen bonds, which greatly reduce its negative charge potential. Similarly to trypsin D189G/G226D, the native Asp²²⁶ of factor B forms hydrogen bonds and this correlates with the low binding affinity and overall low catalytic efficiency toward P₁-Arg/Lys peptide substrates (12–14). Re-constructing the pocket of factor B in the D226N/N189D mutant caused complete loss of hemolytic and C3 proteolytic activity (Figs. 5 and 6), although esterolytic activity toward the P₁-Arg thioester substrate was partially retained (Fig. 7). The kinetic analysis showed that the 80% reduction in esterolytic activity (k_{cat}/K_m) was almost entirely due to reduction in k_{cat} , whereas the K_m was not affected. Thus, the exact location of the negative charge at base of the S₁ site and particularly its spatial relationship to the His⁵⁷-Ser¹⁹⁵ dyad and the oxyanion hole, which is altered in trypsin D189G/G226D and factor B D226N/N189D, are especially critical for efficient catalysis.

In an effort to directly compare factor B to trypsin, a Gly residue was substituted at position 226 either alone (D226G) or in combination with the N189D mutation (D226G/N189D). Neither mutant had hemolytic activity. However, loss of hemolytic activity could not be attributed exclusively to defective substrate recognition at the S₁ site because the ability of these mutants to participate in the assembly of the C3 convertase was also affected (Figs. 3 and 5). Binding of the mutants to CoVF and their sensitivity to factor D cleavage was substantially decreased indicating conformational changes near or at the C3b/CoVF-binding sites, which are presumed to be distal to the mutation sites. Because overall folding of the polypeptide chain and the conformation of antigenic epitopes appeared unaffected, the conformational alteration of the C3b-binding site must be subtle, albeit functionally significant. At present it is not clear how the catalytic center relates spatially to the C3b/CoVF-binding sites. Hourcade *et al.* (30) also described a conformational change at a site distal from the mutation in the F227A mutant (30). The mutant was cleavable by factor D, but cleavage did not promote the conformational change to a high affinity C3b-binding proteolytically active state, which characterizes wt factor B. The Bb fragment of this mutant was recognized by a Bb-specific mAb at much lower efficiency than the wt counterpart. As viewed in the structure of B-SP, the RDFHIN^{225–230} segment forms an extended internal β -strand, which is buried within the protein core. Substituting Ala for Phe at position 227 might destabilize the core, affecting the conformation of the surface epitope recognized by the Bb-specific mAb (30). This epitope is probably located near the RDFHIN^{225–230} segment and is only reactive in Bb perhaps because it is sterically hindered by the Ba region of intact factor B or because it undergoes a conformational change upon cleavage/removal of Ba. Our D226G mutants might have conformational change(s) within the same region. However, the relationship between the possible conformational change of the antigenic epitope and that of the C3b-binding site is still unclear.

It is of interest that the RDFHIN^{225–230} motif is found in factor B and C2 of most animal species, but is absent from all other complement enzymes (1) as well as from other SPs of the

large chymotrypsin family (40, 38). This underlines the fundamental role of Asp²²⁶ in the function of factor B and C2 in complement activation. Therefore, the native conformation of Asp²²⁶ and Asn¹⁸⁹ or Ser¹⁸⁹ within the S₁ pocket of factor B and C2, respectively, constitutes one of the structural determinants, which have evolved to optimize the highly specific C3/C5 cleavage. However, C3/C5 recognition and hydrolysis require more extensive enzyme-substrate contacts than interaction of the side chain of P₁-Arg with residues of the S₁ site. The disparity in catalytic activity toward C3 and dipeptide substrates of N189D and D226N/N189D factor B (Figs. 6 and 7) probably reflects the complexity of the interaction between C3b-bound Bb and its natural substrates, C3 and C5.

In the present study, we correlated the crystal structure of B-SP to the detailed mutational analysis of the factor B S₁ pocket. The resulting information contributes to current understanding of the structural basis for factor B and C2 substrate specificity and catalysis. Such knowledge is crucial for designing highly specific inhibitors that could have therapeutic potential for complement-mediated human diseases.

Acknowledgments—We express our appreciation to Xiao Ying Liu and Yuling Dai for excellent technical assistance.

REFERENCES

- Arlaud, G. J., Volanakis, J. E., Thielens, N. M., Narayana, S. V. L., Rossi, V., and Xu, Y. (1998) *Adv. Immunol.* **69**, 249–307.
- Volanakis, J. E. (1998) in *The Human Complement System in Health and Disease* (Volanakis, J. E., and Frank, M. M., eds) pp. 9–32, Marcel Dekker, Inc., New York.
- Cohen, G. H., Silverton, E. W., and Davies, D. R. (1981) *J. Mol. Biol.* **148**, 449–479.
- Gráf, L., Craik, C. S., Pathy, A., Rocznik, S., Fletterick, R. J., and Rutter, W. J. (1987) *Biochemistry* **26**, 2616–2623.
- Gráf, L., Jancsó, A., Szilágyi, L., Hegyi, G., Pintér, K., Náray-Szabó, G., Hepp, J., Medzihradsky, K., and Rutter, W. J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85**, 4961–4965.
- Schellenberger, V., Braune, K., Hofmann, H.-J., and Jakubke, H.-D. (1991) *Eur. J. Biochem.* **199**, 623–636.
- Craik, C. S., Largman, C., Fletcher, T., Rocznik, S., Barr, P. J., Fletterick, R., and Rutter, W. J. (1985) *Science* **228**, 291–297.
- Jing, H., Macon, K. J., Moore, D., Lawrence, J. D., Volanakis, J. E., and Narayana, S. V. L. (1999) *EMBO J.* **18**, 804–814.
- Volanakis, J. E. (1989) in *Year Immunol. Cellular, Molecular and Clinic Aspects* (Cruse, J. M., and Lewis, R. E., Jr., eds) Vol. 4, pp. 218–230, S. Karger, Basel.
- Pangburn, M. K., and Müller-Eberhard, H. J. (1986) *Biochem. J.* **235**, 723–730.
- Cooper, N. R. (1975) *Biochemistry* **14**, 4245–4251.
- Ikari, N., Hitomi, Y., Nünobe, M., and Fujii, S. (1983) *Biochim. Biophys. Acta* **742**, 318–323.
- Caporale, L. H., Gaber, S.-S., Kell, W., and Götze, O. (1981) *J. Immunol.* **126**, 1963–1965.
- Kam, C.-M., McRae, B. J., Harper, J. W., Niemann, M. A., Volanakis, J. E., and Powers, J. C. (1987) *J. Biol. Chem.* **262**, 3444–3451.
- Horiuchi, T., Kim, S., Matsumoto, M., Watanabe, I., Fujita, S., and Volanakis, J. E. (1993) *Mol. Immunol.* **30**, 1587–1592.
- Zoller, M. J., and Smith, M. (1983) *Methods Enzymol.* **100**, 468–500.
- Kunkel, T. A. (1985) *Proc. Natl. Acad. Sci. U. S. A.* **82**, 488–492.
- Tabor, S., and Richardson, C. C. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84**, 4767–4771.
- Itakura, K., Rossi, J. J., and Wallace, R. B. (1984) *Annu. Rev. Biochem.* **53**, 323–356.
- Agrawal, A., Xu, Y., Ansardi, P., Macon, K. J., and Volanakis, J. E. (1992) *J. Biol. Chem.* **267**, 25353–25358.
- Kim, S., Narayana, S. V. L., and Volanakis, J. E. (1994) *Biochemistry* **33**, 14393–14399.
- Ueda, A., Kearney, J. F., Roux, K. H., and Volanakis, J. E. (1987) *J. Immunol.* **138**, 1143–1149.
- Tuckwell, D. S., Xu, Y., Newham, P., Humphries, M. J., and Volanakis, J. E. (1997) *Biochemistry* **36**, 6605–6613.
- Circolo, A., Nutter, T. B., and Strunk, R. C. (1997) in *Complement, a Practical Approach* (Dodds, A. W., and Sim, R. B., eds) pp. 199–221, Oxford University Press, Oxford.
- Kulics, J., Circolo, A., Strunk, R. C., and Colten, H. R. (1994) *Immunology* **82**, 509–515.
- Volanakis, J. E., and Macon, K. J. (1987) *Anal. Biochem.* **163**, 242–246.
- Gresham, H. D., Matthews, D. F., and Griffin, F. M., Jr. (1986) *Anal. Biochem.* **154**, 454–459.
- Rapp, H. J., and Borsos, T. (1970) in *Molecular basis of Complement Action* (Rapp, H. J., and Borsos, T., eds) Century Crofts, New York.
- Xu, Y., and Volanakis, J. E. (1997) *J. Immunol.* **158**, 5958–5965.
- Houcade, D. E., Mitchell, L. M., and Oglesby, T. J. (1998) *J. Biol. Chem.* **273**, 25996–26000.
- Evnin, L. B., Vásquez, J. R., and Craik, C. S. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87**, 6659–6663.
- Tsu, C. A., Perona, J. J., Schellenberger, V., Truck, C. W., and Craik, C. S. (1994) *J. Biol. Chem.* **269**, 19565–19572.
- Hof, P., Mayr, I., Huber, R., Korzus, E., Potempa, J., Travis, J., Powers, J. C., and Bode, W. (1996) *EMBO J.* **15**, 5481–5491.
- Fujinaga, M., Chernaia, M. M., Halenbeck, R., Kothe, K., and James, M. N. G. (1996) *J. Mol. Biol.* **261**, 267–278.
- Navia, M. A., McKeever, B. M., Springer, J. P., Lin, T.-Y., Williams, H. R., Fluder, E. M., Dorn, C. P., and Hoogsteen, K. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 7–11.
- Polanowska, J., Krokoszynska, I., Czapinska, H., Watorek, W., Dadlez, M., and Otlewski, J. (1998) *Biochim. Biophys. Acta* **1386**, 189–198.
- Tsu, C. A., Perona, J. J., Fletterick, R. J., and Craik, C. S. (1997) *Biochemistry* **36**, 5393–5401.
- Czapinska, H., and Otlewski, J. (1999) *Eur. J. Biochem.* **260**, 571–595.
- Perona, J. J., Tsu, C. A., McGrath, M. E., Craik, C. S., and Fletterick, R. J. (1993) *J. Mol. Biol.* **230**, 934–949.
- Greer, J. (1990) *Proteins: Struct. Funct. Genet.* **7**, 317–334.



Exhibit 42

Complementary DNA Cloning and Sequencing of Rat Enteropeptidase and Tissue Distribution of Its mRNA

Naohisa Yahagi,* Masao Ichinose,*¹ Masashi Matsushima,* Yasuo Matsubara,*
Kazumasa Miki,* Kiyoshi Kurokawa,* Hiroshi Fukamachi,† Kosuke Tashiro,†
Koichiro Shiokawa,† Takeshi Kageyama,‡ Takayuki Takahashi,§ Hideshi Inoue,¶ and
Kenji Takahashi¶

*First Department of Internal Medicine, Faculty of Medicine and †Zoological Institute, Faculty of Science, University
of Tokyo, Tokyo 113, Japan; ‡Department of Molecular and Cellular Biology, Primate Research Institute, Kyoto
University, Inuyama 484, Japan; §Division of Biological Science, Graduate School of Science, Hokkaido University,
Sapporo 060, Japan; and ¶Laboratory of Molecular Biochemistry, Molecular Science Division, School of Life Science,
Tokyo University of Pharmacy and Life Science, Tokyo 192, Japan

Received January 19, 1996

A cDNA clone encoding enteropeptidase (EC 3.4.21.9), a key enzyme for the conversion of trypsinogen to trypsin, was isolated from a rat duodenal mucosa cDNA library. Sequence of the 3585 base pair clone predicted that enteropeptidase is synthesized as a single-chain precursor form, proenteropeptidase, consisting of 1058 amino acid residues with an internal signal sequence (51 residues) and is then processed into the mature enzyme consisting of three different peptide chains, i.e., mini, light and heavy chains, not the previously reported two-chain enzyme. The structure of enteropeptidase is relatively conserved among different species and the rat enteropeptidase is 24 and 39 amino acids longer than the porcine and human ones, respectively. Northern blot analysis of RNAs from normal rat tissues revealed that the enteropeptidase mRNA of around 4.4 kb in size was expressed only in the duodenal mucosa, and high proteolytic activity of the enzyme was detected in the proximal small intestine. Additional analysis of the RNAs by RT-PCR revealed that a low level of the mRNA was also expressed in the other parts of the small intestine, i.e., jejunum and ileum. These results indicate that the biosynthesis of enteropeptidase takes place mainly in the proximal small intestine, the duodenum, and the importance of the region in the physiology of intestinal protein digestion regulated by the enzyme is suggested. Furthermore a faint signal of the mRNA was also detected in the stomach, colon and brain in which the existence of trypsin-like serine proteases were reported. The significance of the low level expression of the gene is unclear, but the potential peptide-processing function of the enzyme in these tissues is also suggested. © 1996 Academic Press, Inc.

Enteropeptidase (Enterokinase EC 3.4.21.9) was initially recognized as an intestinal factor which activates the latent enzymes in pancreatic fluid. Later the enzyme was proved to be involved in the conversion of trypsinogen to trypsin (1), leading to the activation of various pancreatic zymogens involved in the later stages of the digestive cascade. Therefore, enteropeptidase has been considered to be a key enzyme in the intestinal protein digestion. Because of its medical and physiological importance, the enzyme has been purified from the small intestine of various species, including bovine (2), porcine (3) and human (4). In addition, their cDNA structure have recently been determined in these species by us and others (5-8). However, the details of the structure and function of the enzyme are still unclear now. For example, the number of the peptide chains composing the mature enzyme is differently reported depending on the species and the mechanism of the enzyme activation remains to be elucidated. Also unclear is the regulatory mechanism of its synthesis in the gastrointestinal tract. In order to clarify these problems and because the laboratory rat is a highly developed experimental model to study the physiology of the intestinal digestion, we attempted to characterize rat enteropeptidase. In this study, we determined the nucleotide sequence

¹ Corresponding author. Fax: 03-3812-5063.

Tissue prepar
Wistar strain m
phosphate buffe
the guanidium
oligo(dT)-cellul
fluorometrically
naphthylamide).
Isolation and
for the preparati
Hoffman (10).
according to the
of lambda ZAP
coli strain XL-1
enteropeptidase
purified phages
manufacturer an
method on both
Inch, a thermal
mRNA detecti
subjected to ele
membrane filter
stringency condi
of ADNA genera
(5' primer, 5'-A
ment, 491bp)
TAGGCCATGA
and purified by
to cDNA and th
Jerman) under the
conditions, the a
PCR products w
staining.

Approxima
teropeptidase
clones, 50 clo

P
5'
T

FIG. 1. Restr
constructed from
described in Mai
pBluescript. Line
proenteropeptida:

of cDNA encoding rat enteropeptidase, predicted primary structure of the enzyme, and analyzed the gene expression in the rat digestive tract and various other organs.

MATERIALS AND METHODS

Tissue preparation, RNA isolation and assay of enzymatic activities of enteropeptidase. All tissues were collected from Wistar strain male adult rats (8 weeks old, Charles River Japan, Inc.). The excised tissues were washed with ice-cold phosphate buffered saline and were stored frozen in liquid nitrogen until use. From the tissues, total RNA was prepared by the guanidium isothiocyanate/cesium chloride density gradient ultracentrifugation and poly(A)⁺RNA was selected by oligo(dT)-cellulose column chromatography. The proteolytic activity of enteropeptidase in the tissue samples was measured fluorometrically by a modified method of Antonowicz et al. (9), using a synthetic substrate [Gly-(Asp)₄-Lys-β-naphthylamide]. Unless otherwise specified, 2mM EDTA was included in the reaction mixture in this study.

Isolation and characterization of the cDNA clone for rat enteropeptidase. Rat duodenal mucosa poly(A)⁺ RNA was used for the preparation of a cDNA library. Double-stranded cDNA was synthesized according to the procedure of Gubler and Hoffman (10). After methylation of the internal EcoRI sites and addition of EcoRI linkers, the cDNAs were fractionated according to their size by agarose-gel electrophoresis. The cDNA larger than 1.5kb in length was ligated into the EcoRI sites of lambda ZAP II vector (Stratagene, USA). The phages were packaged and recombinants were selected by plating on *E. coli* strain XL-1 blue. Nylon filters that carried denatured recombinant DNAs were screened by [³²P]-labeled porcine enteropeptidase cDNA (7). The positively hybridized clones were identified and isolated by repeated purification. The purified phages were converted to the corresponding plasmid form utilizing the plasmid excision procedure provided by the manufacturer and were used as a template for DNA sequencing. Sequencing was performed by dideoxy chain termination method on both strands of denatured plasmid cDNA inserts using a Taq dye terminator sequencing kit (Applied Biosystems, Inc.), a thermal cycler (model 480, Perkin Elmer Cetus), and a DNA sequencer (model 371A, Applied Biosystems, Inc.).

RNA detection by Northern blotting and RT-PCR. 10 μg of total RNA from various rat tissues were denatured and subjected to electrophoresis on a 0.66 M formaldehyde-agarose gel. After the RNA had been transferred to a nylon membrane filter, the filter was hybridized with the [³²P]-labeled full-length cDNA for rat enteropeptidase under high-stringency conditions. The size of RNA was estimated by reference to the mobility of 18s and 28s rRNAs and fragments of λDNA generated by digestion with Hind III. Primers specific for the amplification of the rat enteropeptidase heavy chain (5' primer, 5'-ATTGATGATGGTTTGG-3'; 3' primer 5'-AGGTTGGTCTGGATAAG-3'; size of the amplified fragment, 491bp) and G3PDH (5' primer, 5'-TGAAGGTCGGTGTCAACGGATTGGC-3'; 3' primer 5'-CATGTAGGCCATGAGGTCCACCAC-3') were synthesized with a DNA synthesizer (model 380A, Applied Biosystems, Inc.) and purified by gel filtration. For each reaction, 1 μg of poly(A)⁺ RNA from representative tissues was reverse-transcribed to cDNA and the resulting cDNA was subjected to 20 to 40 cycles of PCR using Takara Taq DNA polymerase (Takara, Japan) under the following conditions: 94°C for 60sec. → 48°C for 30sec. → 74°C for 60sec. In the above-mentioned conditions, the amplified signal derived from the genomic DNA encoding enteropeptidase was around 1.6kb in size. The PCR products were electrophoresed through a 1.0% agarose gel in IX TAE buffer and visualized by ethidium bromide staining.

RESULTS AND DISCUSSION

Approximately 5×10^5 clones were screened by hybridization with a full-length porcine enteropeptidase cDNA. Over 500 clones were identified as positive for the probe. Among these clones, 50 clones hybridized positively with 0.6 kb EcoRV fragment representing the NH₂-terminal

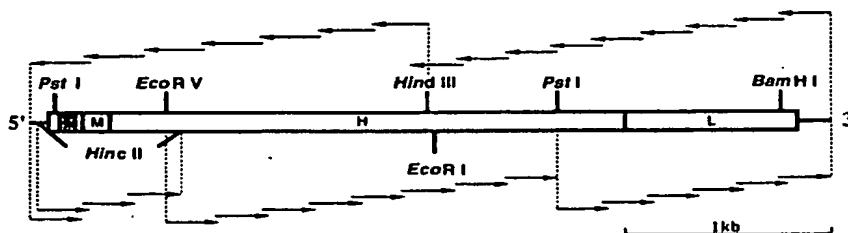


FIG. 1. Restriction map and sequencing strategy of a rat enteropeptidase cDNA clone (REK#7). Deletion mutants constructed from subcloned fragments were used for nucleotide sequencing, and sequencing was done in both directions as described in Materials and Methods. Arrows indicate the direction and extent of sequencing of fragments subcloned in pBluescript. Lines indicate the 5'- and 3'-noncoding region, a closed box indicates the putative internal signal sequence of proenteropeptidase. Open boxes indicate the coding region including the M, H and L-chains of mature enteropeptidase.

domain of porcine enteropeptidase. These clones were isolated by repeated purification. The restriction site map constructed for these clones revealed that their structures are basically the same and the nucleotide sequencing on bilateral ends disclosed a common nucleotide sequence. One clone (REK#7) was found to contain the entire coding region for rat enteropeptidase. The restriction map and the sequence analysis strategy of the clone is shown in Fig. 1. The resulting nucleotide sequence and the deduced amino acid sequence of rat enteropeptidase are presented in Fig. 2. The analyzed cDNA clone was 3585 base pairs (bp) long, including the 5'-noncoding region (166bp), the coding nucleotide sequence (3,174bp) and the 3'-noncoding region (245bp). A typical polyadenylation signal was present at the 3554th base pair position. The second methionine codon at nucleotides 167-169 in the open reading frame meets the criteria for the initiation site of the translation (11). Thus, the cDNA encoding rat enteropeptidase predicts a molecule of 1058 amino acids residues ($M_r = 117,700$). Recently, we purified the enzyme from porcine duodenal mucosa and structurally characterized it. In addition, we have cloned and analyzed the cDNA coding for the protein (7). The primary structures of the rat and porcine enzymes are relatively conserved; 77% identical in the nucleotide sequence and 71% in the encoded amino acid sequence. The comparison of the rat cDNA sequence with the porcine one indicated that the enzyme is originally synthesized as a single-chain precursor and processed into a three-chain enzyme rather than the heterodimeric enzyme previously reported in other species (2,3). The NH_2 -terminal sequences of the mini (M), heavy (H), and light (L)-chains are deduced to start at positions 53, 119, and 819, respectively, thus leading to the production of three chains consisting of 66 ($M_r = 7,700$), 700 ($M_r = 77,700$), and 240 ($M_r = 26,800$) amino acid residues. There is a hydrophobic domain comprising 25 amino acid residues preceding the NH_2 -terminus in the rat proenteropeptidase sequence; double underlined region from positions 19 to 43. Although there is one amino acid insertion (Ala at position 52) in the prepeptide sequence compared with other species (6,7), the hydrophobic segment is observed in common, probably serving as an internal signal sequence. While we were preparing the manuscript, the sequence of the cDNA encoding human enteropeptidase was reported, presenting the possibility of a two-chain structure of the human enzyme (8). However, it is noteworthy that in addition to the H and L-chains, a sequence similar to the rat and porcine M-chains is also observed in the human sequence. The homology of the region is particularly high (88% vs. porcine and 83% vs. human enzymes) compared with that in other regions (64-68% in the H-chain, 77-78% in the L-chain). Thus, it is highly probable that human enteropeptidase is also a three-chain enzyme. Among these three chains, the homology of the H-chain is the lowest due to insertions/deletions of variable length around the Ser/Thr-rich regions, potential O-linked glycosylation sites. The rat enzyme has 7 insertions (18 amino acids in total) and 50% of the inserted amino acids are Ser and Thr residues, which are probably involved in O-linked carbohydrate attachment. The rat enzyme is therefore considered to be the most O-linked carbohydrate-rich enteropeptidase among the previously reported species. Furthermore, some of these inserted amino acids give rise to two additional potential N-glycosylation sites, leading to heavy glycosylation of the region. The number of potential N-linked glycosylation sites is variable depending on species (rat 20, human 18, bovine 19 and porcine 22), but their positions are almost conserved. These carbohydrate moieties are presumably important to protect the enzyme from the access of other digestive proteases in the intestinal content. The variety of the glycosylation sites observed among species may somehow be related with the divergence in the environment in the intestinal lumen and physiology of digestion. The common basic structure of the catalytic domain of serine proteases is also observed in the rat

FIG. 2. Nucleotide and deduced amino acid sequences of the rat enteropeptidase cDNA clone. Double underlined sequence indicates a putative internal signal sequence. Boxed domains with (M), (H) and (L) are the deduced regions, corresponding to the M, H and L-chains of the mature enzyme, respectively. The underlined sequence at 665-802bp is the variable and Ser/Thr rich region, including 18 amino acid residues of insertions observed in the rat enzyme. Potential N-linked glycosylation sites are indicated by closed boxes.

	C	CC
93	TAT TCG	
94	Ala Gly	
193	GCA GGA	
41	Leu Ala	
197	CTG GCA	
39	TTC AAG	
203	CCG AAT	
169	Acc Gln	
679	AAT GAA	
137	Val Gln	
179	GTG AAA	
129	Leu Ser	
171	TGG AGC	
201	ACT TTT	
177	ACC GCA	
223	Val Thr	
663	GTA ACT	
269	Leu Thr	
189	TTC ACG	
297	Leu Ser	
1055	CTT TCC	
329	Ser Phe	
1551	TCT TTT	
361	Phe Thr	
1247	TTT AAG	
11	Arg Leu	
111	GAT CTT	
433	Gly Phe	
1639	GGA TTT	
457	Leu His	
1539	ACT CAY	
489	Gln Gln	
1671	GAA GAA	
521	Cys Ser	
1757	TGC AGT	
593	Gln His	
1893	GAG CAY	
593	Phe Cys	
1911	TTC TGT	
211	Arg Asp	
2010	AGA GAT	
649	Ile Phe	
2333	ATT TTC	
693	Gln Phe	
2307	GAA TTT	
713	Arg Phe	
1303	CCT TTT	
745	Ile Ser	
2399	ATT TCA	
777	Gln Ala	
1695	GAA GCT	
799	Lys Met	
2111	AAG ATG	
111	Arg Ser	
2007	AGG TCT	
673	Thr Arg	
2703	ACC AGA	
803	His Tyr	
2679	CAT TAT	
237	Gln Gln	
2379	GAA GAA	
819	Ala Asp	
2071	GCT GAT	
1001	Gly Thr	
2167	GCG ACA	
1111	Cys Ala	
2111	TCT GCA	
2359	CAT AAA	
2655	CTA AAT	
2651	CAT AAT	

cDNA clone. Double underlined and (L) are the deduced regions. The underlined sequence at 665–802bp is the deduced sequence in the rat enzyme. Potential

[illegible]

L-chain. Consistent with the previous data indicating that the enzyme activity is attained by the L-chain alone, the homology of the region is high among different species (77–78%). There is, however, an insertion of 4 amino acid residues in the sequence next to the catalytic triad of serine proteinases, whereas the three basic amino acid residues, important to keep the substrate specificity for the trypsinogen, are well conserved.

The expression of the enteropeptidase gene in various rat tissues was examined by Northern blot analysis using the cloned full-length cDNA as a probe. As shown in Fig.3, a signal of 4.4 kb enteropeptidase mRNA was observed only in the duodenum, but not in the other parts of gastrointestinal tract from the esophagus to the colon and also not in other organs such as the brain, heart, lung, liver, kidney and spleen. Since the comparable signal for G3PDH mRNA was observed in all RNA samples analyzed, it is evident that the paucity of the enteropeptidase mRNA in the jejunum and ileum was not caused by the degradation of the RNAs. There is a controversy as for the distribution of enteropeptidase; some of the previous reports indicated the limited localization of the enzyme in the duodenum (12), while others the distribution throughout the small intestine (9). Thus, to further measure low levels of enteropeptidase gene expression semiquantitatively, we employed the RT-PCR method and selected a primer set and amplification conditions with high sensitivity and low background. Three PCR cycles were used for quantitative estimation. The RT-PCR result of the RNA samples used in the Northern blotting is shown in Fig.5. The PCR product had a molecular size of 0.5 kb corresponding to the expected product of 491 bp and was shown to hybridize with the rat enteropeptidase cDNA by Southern blotting (data not shown). A strong signal was observed in the duodenum and also weak signals in the jejunum and ileum. The signal detected in the ileum at 34 cycles was weaker than that of the duodenum at 30 cycles. Thus, the mRNA level in the duodenum is considered to be at least 10 times higher than that in the distal part of small intestine, the ileum end. These results indicate the gene expression of the enzyme along the entire small intestine, though the level of the expression is low in the distal segment. Previous studies revealed relatively high enzyme activity throughout rat small intestine (9). The analysis of our samples by the same assay for the enzyme activity also gave essentially the same result (Fig.4/A). However, it was indicated that their method also measured the coexisting aminopeptidase activity together (13). By including 2mM EDTA in the reaction buffer, the activity of aminopeptidase could be completely diminished, while that of enteropeptidase was not much affected, at least 80% of the activity having remained (unpublished data). Thus, an approximate estimate for the enteropeptidase level could be obtained by the method used in the presence of

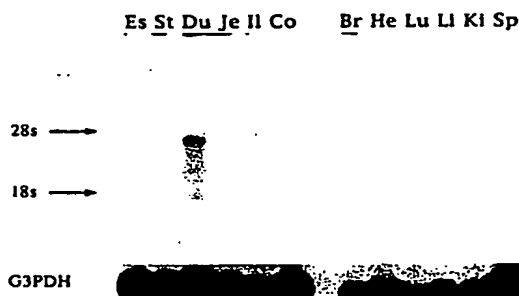


FIG. 3. Northern blot analysis of total cellular RNA from various rat organs. 10 μ g of total cellular RNA from the rat esophagus (Es), stomach (St), duodenum (Du), jejunum (Je), ileum (Il), colon (Co), brain (Br), heart (He), lung (Lu), liver (Li), spleen (Sp) and kidney (Ki) were separated on a 1.0% denaturing/formaldehyde agarose gel, hybridized and washed under the condition of high stringency using the rat enteropeptidase cDNA as a probe. The lines on the left indicate the positions of the 28S and 18S ribosomal RNAs. The results of rehybridization of the filter with glyceraldehyde-3-phosphate dehydrogenase (G3PDH) cDNA are shown at the bottom.

FIG. 4. Enter divided into 8 eq section (A: with percentage of the

2mM EDTA indicates the p no enzyme ac Taken together dase is regula place of the pancreatic sec

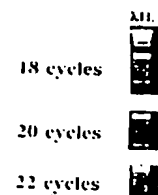


FIG. 5. Enter (Du), jejunum (Je) adjusted to the s confirm the expo enteropeptidase E

y is attained by the 77-78%). There is, lytic triad of serine substrate specificity

ed by Northern blot a signal of 4.4 kb her parts of gastro- as the brain, heart, was observed in all tNA in the jejunum troversy as for the ited localization of small intestine (9). iquantitatively, we onditions with high ve estimation. The in Fig.5. The PCR of 491 bp and was data not shown). A um and ileum. The at 30 cycles. Thus, an that in the distal ion of the enzyme the distal segment. l intestine (9). The ssentially the same he coexisting ami- ffer, the activity of ase was not much us, an approximate in the presence of

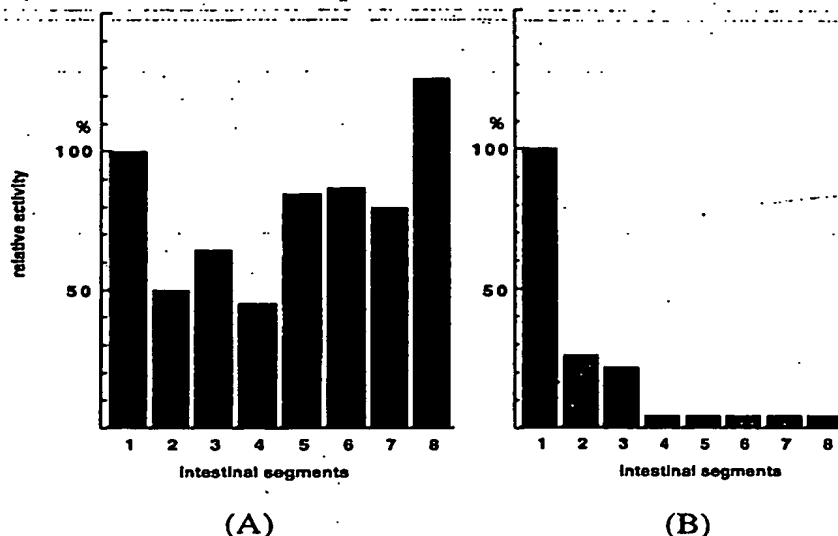


FIG. 4. Enteropeptidase activity along the rat small intestine. Small intestine from the duodenum to the ileum end was divided into 8 equal segments, and the activity in each segment was measured as described in the Materials and Methods section (A: without EDTA, B: with 2mM EDTA in the reaction, respectively). Value of each segment indicates the percentage of the enzyme activity when that in the duodenum (segment No. 1) is regarded as 100%.

2mM EDTA and the result of the measurements in the small intestine is shown in Fig.4/B. This indicates the presence of high enzyme activity in the proximal segment of the small intestine, while enzyme activity was detected in the distal segment despite the high sensitivity of the method. Taken together, the above-mentioned results clearly indicate that the biosynthesis of enteropeptidase is regulated region-specifically both at the level of transcription and translation and that main place of the synthesis is the proximal segment of the small intestine, where pancreatic secretion join the intestinal contents. The distribution of the mRNA and the enteropep-

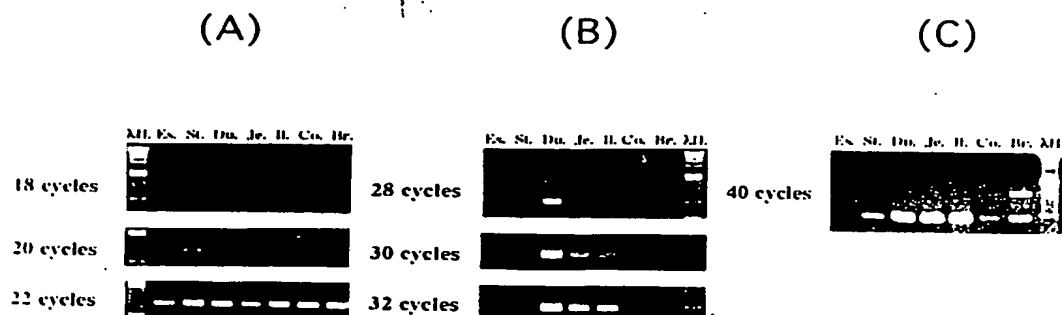


FIG. 5. Enteropeptidase mRNA expression detected by RT-PCR in the rat esophagus (Es), stomach (St), duodenum (Du), jejunum (Je), ileum (Il), colon (Co) and brain (Br). The amount of each cDNA sample included in the reaction was adjusted to the same quantity according to the G3PDH mRNA expression. Three successive cycles were employed to confirm the exponential amplification. Primers used for the amplification were as follows: (A) G3PDH, (B) and (C) rat enteropeptidase H-chain specific primer.

tidase activity strongly indicate the importance of the proximal small intestine in the physiology of the intestinal protein digestion regulated by the enzyme.

In addition, faint signals of enteropeptidase mRNA were also observed in the stomach, colon and brain at 40 cycles (Fig.5/C). The enzyme activity is undetectable in these organs and the physiological importance of these findings remain to be elucidated. However, these findings are interesting in context with the previous reports indicating the presence of trypsin-like serine proteases in these tissues (14, 15). Trypsin-like serine proteases are playing important roles in many biological processes. Especially in human brain, they are considered to be involved in the pathogenesis of Alzheimer's disease, playing a role in β -amyloid production (15). Thus, the observed distribution of the mRNA may indicate a role of enteropeptidase in the processing of bioactive peptides by regulating the activity of trypsin-like proteases.

REFERENCES

1. Kunitz, M. (1939) *J. Gen. Physiol.* 22, 429-446.
2. Liepnieks, J. J., and Light, A. (1979) *J. Biol. Chem.* 254, 1677-1683.
3. Baratti, J., Maroux, S., Louvard, D., and Desnuelle, P. (1973) *Biochim. Biophys. Acta* 315, 147-161.
4. Magee, A. I., Grant, D. A. W., and Hermon-Taylor, J. (1981) *Clin. Chim. Acta* 115, 241-254.
5. LaVallie, E. R., Rehemtulla, A., Racie, L. A., DiBlasio, E. A., Ferenz, C., Grant, K. L., Light, A., and McCoy, J. M. (1993) *J. Biol. Chem.* 268, 23311-23317.
6. Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W., and Sadler, J. E. (1994) *Proc. Natl. Acad. Sci. USA* 91, 7588-7592.
7. Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., Umeyama, H., Inoue, H., Takahashi, T., and Takahashi, K. (1994) *J. Biol. Chem.* 269, 19976-19982.
8. Kitamoto, Y., Veile, R. A., Donis-Keller, H., and Sadler, J. E. (1995) *Biochemistry* 34, 4562-4568.
9. Antonowicz, I., Hesford, F. J., Green, J. R., Grogg, P., and Hardorn, B. (1980) *Clin. Chim. Acta* 101, 69-76.
10. Gubler, V., and Hoffman, B. J. (1983) *Gene* 25, 263-269.
11. Kozak, M. (1984) *Nucleic Acids Res.* 12, 857-872.
12. Eggermont, E., Molla, A. M., Tytgat, G., and Rutgeerts, L. (1971) *Acta Gastro-Enterologica Belgica* 34, 655-662.
13. Jenö, P., Green, J. R., and Lentze, M. J. (1987) *Biochem. J.* 241, 721-727.
14. Jeohn, G. H., Serizawa, S., Iwamatsu, A., and Takahashi, K. (1995) *J. Biol. Chem.* 270, 14748-14755.
15. Wiegand, U., Corbach, S., Minn, A., Kang, J., and Müller-Hill, B. (1993) *Gene* 136, 167-175.

T

Aki

*Faculty
066

Th
this n
the p
acid
prote
the n
kinas
speci
prote
that
repre

The N
is comp
served
bHLH-2
suggest
the maj
(NCAM
c-myc v
proteins
from al
HLA- β
elemen
upstrea
shows
GATT
transcri
gene to
pressio
48 KD
to reco
N-myc
H2TF1
interfe
mecha

To

Exhibit 43

Cloning and Characterization of the cDNA for Human Airway Trypsin-like Protease*

(Received for publication, May 6, 1997, and in revised form, March 2, 1998)

Kazuyoshi Yamaoka†§, Ken-ichi Masuda†, Hiroko Ogawa†, Ken-ichiro Takagi†, Naoji Umemoto†, and Susumu Yasuoka¶

From the †Teijin Institute for Biomedical Research, 4-3-2 Asahigaoka, Hino, Tokyo 191 and the ¶Department of Nursing, School of Medical Sciences, University of Tokushima, 3-18-15 Kuramotocho, Tokushima City, Tokushima 770, Japan

Previously we isolated a trypsin-like enzyme designated human airway trypsin-like protease from the sputum of patients with chronic airway diseases. This paper describes the cDNA cloning, characterization of the primary protein structure deduced from the cDNA, and gene expression of this enzyme in various human tissues. We obtained an entire 1517-base pair sequence of cDNA with an open reading frame encoding a polypeptide with 418-amino acid residues. The polypeptide consisted of a 232-residue catalytic region and a 186-residue noncatalytic region with a hydrophobic putative transmembrane domain near the NH₂ terminus. The polypeptide was suggested to be a type II integral membrane protein in which the COOH-terminal catalytic region is extracellular. Therefore, this protein is thought to be synthesized as a membrane-bound precursor and to mature to a soluble and active protease by limited proteolysis. It showed 29–38% identity in the sequence of the catalytic region with human hepsin, enteropeptidase, acrosin, and mast cell tryptase. The noncatalytic region had little similarity to other known proteins. In Northern blot analysis a transcript of 1.9 kilobases was detectable most prominently in the trachea among 17 human tissues examined.

Many previous investigations have indicated that proteases released from immunoinflammatory cells participate in pathogenesis of several kinds of respiratory diseases. For instance, neutrophil elastase has been shown to be intimately related to the pathologic states of pulmonary emphysema (1, 2), cystic fibrosis (3, 4), interstitial pneumonia (5), and adult respiratory distress syndrome (6) through destruction of extracellular matrix components, such as elastin, of alveolar and bronchial tissues. Mast cells, which abound in airway mucosa and in alveolar wall, release trypsin-like protease (tryptase) and chymotrypsin-like protease (chymase) into extracellular spaces during degranulation (7). The tryptase has potential to stimulate smooth muscle, fibroblast, and tissue turnover (8). Different substrates for chymase (9–11) indicate the potential involvement of the enzyme in a variety of processes related to the inflammatory response. Recently it was revealed that chymase

from human mast cells selectively converted big endothelins to trachea-constricting peptides (12). These effects of the two mast cell proteases have attracted considerable attention as one of the pathogenic determinants and the therapeutic targets of bronchial asthma and allergic inflammation. Elastase released from alveolar macrophages has also been suggested to contribute to the pathogenesis of pulmonary emphysema by degrading matrix components of alveolar walls (13, 14).

However, there are very few reports dealing with the functions and roles of proteases secreted from respiratory tissues, such as secretory glands or surface epithelial cells of the airway. Kido and co-workers (15, 16) found a novel trypsin-like protease that is secreted from rat Clara cells, secretory cells localized to the distal airway only. The protease, named tryptase Clara, was shown to enhance the infectivity of influenza and Sendai viruses (17), although its physiological role is unknown.

Previously, we found trypsin-like activity in the sputum of patients with chronic airway diseases and isolated a novel trypsin-like protease from the sputum, designated human airway trypsin-like protease (HAT)¹ (18). Gel filtration studies showed that HAT was a monomeric enzyme with an apparent molecular mass of 27 kDa. Immunohistochemical studies showed that HAT was localized mainly in cells of submucosal serous glands of the bronchi and trachea. These results indicate that HAT is released from the submucosal serous glands onto mucous membrane, at least in patients with chronic airway diseases.

In this paper, we report the cloning of HAT cDNA, the primary structure of this enzyme and characterization of the polypeptide deduced from the nucleotide sequence of the cDNA, and results of analysis of expression of HAT mRNA in various human tissues. The primary structure of HAT was compared with that of other known serine proteases.

EXPERIMENTAL PROCEDURES

Materials—Human trachea QUICK-Clone™ cDNA, human trachea poly(A)⁺ RNA, human trachea λgt10 cDNA library (oligo(dT) and random-primed), 5'-RACE kit, human multiple tissue Northern blots, and human β-actin cDNA were purchased from CLONTECH Laboratories Inc. (Palo Alto, CA). Taq DNA polymerase was from Promega Corp. (Madison, WI). SureClone™ ligation kit, dNTP, and plasmid vector pUC18 were from Amersham Pharmacia Biotech. Avian myeloblastosis virus reverse transcriptase and RNase inhibitor were from Boehringer Mannheim. Restriction endonucleases, random primer labeling kit, and *Escherichia coli* JM109 were from Takara Shuzo Co. Ltd. (Otsu, Japan). Nylon membrane Hybond™-N+ for blotting and [³²P]dCTP for probe labeling in hybridization were from Amersham. Denhardt's solution and salmon sperm DNA were from Wako Pure Chemical Industries Ltd. (Osaka, Japan). Qiagen lambda kit for purification of phage DNA was

* This work was supported by Grant-in-aid for Developmental Scientific Research 09670616 of the Ministry of Education, Science, Sports, and Culture, Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AB002134.

§ To whom correspondence should be addressed. Tel.: 81-425-86-8135; Fax: 81-425-87-5516; E-mail: ymo35291@token1.teijin.co.jp.

from Qiagen GmbH (Hilden, Germany). Oligonucleotide purification cartridge column and DyeDeoxyTM terminator cycle sequencing kit for sequencing of DNA were from Applied Biosystems Inc. (Foster City, CA).

DNA Amplification by Polymerase Chain Reaction (PCR)—PCR was performed according to the procedure described by Sambrook *et al.* (19). Oligonucleotides used as PCR primers were synthesized by a DNA/RNA synthesizer (Applied Biosystems Inc., model 394) and purified by oligonucleotide purification cartridge column. Unless otherwise stated, PCR was carried out by adding 15 pmol of each primer and an appropriate amount of template DNA to 20 μ l of PCR buffer (10 mM Tris-HCl, pH 9.0, 50 mM KCl, 1.5 mM MgCl₂, 1% Triton X-100) containing 0.5 units of *Taq* DNA polymerase and 0.2 mM dNTP. The reaction using a DNA thermal cycler (Perkin-Elmer Corp.) was carried out for 35 cycles of 1-min denaturation at 94 °C, 1.5-min annealing at 57 °C, and 2-min extension at 72 °C.

Subcloning of DNA Fragments—To subclone DNA fragments that were amplified by PCR, SureCloneTM ligation kit was used. DNA fragments were bluntly by Klenow fragment, inserted into the *Sma*I site of plasmid vector pUC18, and introduced into *E. coli* JM109 by Hanahan's method (20). On the other hand, for subcloning of insert DNA of λ gt10 phage clone, the insert DNA was excised by *Eco*RI from phage DNA, which was purified using Qiagen lambda kit and inserted into the *Eco*RI site of plasmid vector pUC18. *E. coli* JM109 was transformed as described above. Plasmid DNA was isolated from each transformant by the alkaline lysis procedure (21) with minor modifications.

Analysis of DNA and Amino Acid Sequence—The nucleotide sequence of the DNA inserted into plasmid vector pUC18 was analyzed by an automated DNA sequencer (Applied Biosystems Inc., model 373) using the Dye-DeoxyTM terminator cycle sequencing kit. Both strands of all clones were completely sequenced. Hydrophathy of amino acid sequence was analyzed (22) with the Genetyx program package (Software Development Co. Ltd., Tokyo, Japan). A computer survey of the National Biomedical Research Foundation (Washington, D.C.) and SWISS-PLLOT (European Bioinformatics Institute, Geneva, Switzerland) data banks for similarity of amino acid sequences between HAT and other known proteins was carried out using MPsrch program, which was modified from the method of Smith and Waterman (23) with Teijin Systems Technology Ltd. (Yokohama, Japan).

Amplification of a Partial cDNA Fragment—In a previous report (18), we showed that the sequence of the 20 NH₂-terminal amino acids of native HAT purified from the sputum of patients with chronic airway diseases was ILGGTEAEEGSPWQVSLRL (amino acids 187–206 in Fig. 1). Based on this amino acid sequence, we designed and synthesized two kinds of degenerate PCR primers; namely 5'-ATCTNGGRC-GNACNGAGGC-3'² (sense) and 5'-ARKCKMAGGCTSACTG-3'² (antisense) to obtain the 59-bp cDNA fragment encoding the front 19 residues of the NH₂-terminal amino acid sequence by PCR. PCR was carried out in the reaction mixture containing 5 pmol of each primer and 1 ng of cDNA derived from human trachea (QUICK-CloneTM cDNA). The amplified DNA fragment was then subcloned and sequenced as described above. The analysis of the sequence showed that a 59-bp DNA fragment encoding the 19-residue amino acid sequence corresponding to the NH₂ terminus of the purified HAT was produced by this PCR.

Amplification of cDNA by 3'-Rapid Amplification of cDNA Ends (RACE)—To obtain a cDNA that had a nucleotide sequence in the downstream side of the 59-bp DNA fragment, we employed the 3'-RACE method developed by Frohman *et al.* (24). Two kinds of sense primers were used to amplify the cDNA specifically and effectively. These primers were designed and synthesized based on the nucleotide sequence of the 59-bp cDNA fragment. At first, single-stranded cDNAs were synthesized by reverse transcription at 42 °C for 60 min in 20 μ l of reaction buffer (50 mM Tris-HCl, pH 7.6, 60 mM KCl, 10 mM MgCl₂, 1 mM dithiothreitol) containing 10 ng of human trachea poly(A)⁺ RNA, 115 pmol of (dT)₁₇-adapter primer 5'-GACTCGAGTCGACATCGA(dT)₁₇-3', 25 units of RNase inhibitor, 1 mM dNTP, and 40 units of avian myeloblastosis virus reverse transcriptase. One-tenth of the reaction mixture was used as a template in the first-round PCR in which 5'-ATCTTGGGGGCGACGGAGGCTGA-3' and the adapter primer 5'-GACTCGAGTCGACATCGAT-3' were used as the sense and antisense primers, respectively. For further amplification of the cDNA, the second-round PCR was carried out using one-fortieth of the first-round PCR reaction mixture as the template with 5'-GAGGCTGAGGAGGGAAGCTGGC-

CGT-3' (nucleotides 635–659 in Fig. 1) and the (dT)₁₇-adapter primer described above as the sense and antisense primers, respectively. The cDNA amplified by 3'-RACE was then subcloned and sequenced.

Screening of cDNA Library—Plaque hybridization against human trachea cDNA library was performed according to the standard procedure (19). The DNA fragment obtained by 3'-RACE was labeled by the random prime method (25) using [α -³²P]dCTP and random primer labeling kit. Using this probe, 1 \times 10⁶ plaques derived from human trachea λ gt10 cDNA library were screened by hybridization as follows. The blots for the plaques were hybridized with the probe at 65 °C overnight (16–20 h) in a solution containing 5 \times SSPE buffer (0.75 M NaCl, 50 mM NaH₂PO₄, 5 mM EDTA, pH 7.4), 5 \times Denhardt's solution, 0.1% SDS, and 100 μ g/ml denatured salmon sperm DNA. These blots were then washed twice at 65 °C for 20 min with 0.1 \times SSPE buffer containing 0.1% SDS. Five positive clones were selected and plaque-purified, and the insert DNAs of these clones were then subcloned and sequenced.

Amplification of cDNA by 5'-RACE—To obtain a cDNA that had a nucleotide sequence in the upstream side of the cDNA coding for native HAT, amplification of the cDNA was carried out using 5'-RACE kit (24). Single-stranded cDNAs were synthesized by reverse transcription of 2 μ g of human trachea poly(A)⁺ RNA using the antisense primer 5'-ACGTGGCAATCCAGTCACGAGGATT-3' (nucleotides 785–761 in Fig. 1). The single-stranded cDNAs were purified using glass powder in 5'-RACE kit after alkaline hydrolysis of RNA in the reaction mixture. Using T4 RNA ligase, AmpliFINDERTM anchor was ligated to the 3'-ends of the single-stranded cDNAs. PCR amplification (0.75 min at 94 °C, 0.75 min at 57 °C, and 2 min at 72 °C) was then carried out using 0.01 of the ligation mixture as template, with anchor primer 5'-CTGGTTCGGCCCCACCTCTGAAGGTTCCAGATCGATAG-3' and 5'-TGA-GCTGCTGTCAGGATCCACATGT-3' (nucleotides 741–717 in Fig. 1) as the sense and antisense primers, respectively. The cDNA amplified by 5'-RACE was then subcloned and sequenced.

Expression and Purification of Recombinant HAT—A 1.3-kb *Bam*HI-*Hind*III fragment containing the entire HAT cDNA was cloned into the transfer vector pBlueBacIII (Invitrogen, San Diego, CA) to generate pBacPHAT1. Recombinant HAT-expressing viruses were generated after co-transfection of Sf9 cells with pBacPHAT1 and wild-type AcMNPV DNA essentially as described by the manufacturer (Invitrogen). For baculovirus/insect cell expression (26), 800 ml of Tn5 (27) cells were then infected with the high titer lysate for 72 h and harvested by centrifugation. The cell pellet was treated with 1% Triton X-100 for 1 h on ice and was centrifuged at 100,000 \times g for 1 h at 4 °C. From this infected cell lysate, the recombinant HAT was isolated by sequential chromatographic procedures of the native HAT purification described previously (18). SDS-polyacrylamide gel electrophoresis, immunoblotting, and degradation of fibrinogen by HAT were done as described (18).

Northern Blot Analysis—The expression level of HAT mRNA in various human tissues was examined by Northern blot analysis. To prepare the probe for the analysis, the full-length cDNA for HAT was ³²P-labeled by random priming (25) and hybridized as follows. Northern blots of various human tissues, which contained 2 μ g of poly(A)⁺ RNA derived from various tissues in each lane, were probed under the same conditions as the library screening described above (except that the concentration of SDS was 0.5%) and then washed. In the case of the blot for trachea, 2 μ g of human trachea poly(A)⁺ RNA was resolved by 1% agarose-formaldehyde gel electrophoresis (28), and transferred onto HybondTM-N+ blotting membrane and UV-cross-linked. X-ray films were exposed to the probed blots for 4 days at -80 °C with an intensifying screen, and the presence of HAT mRNA in each human tissue was evaluated. These blots were then stripped of the HAT cDNA probe by boiling in 0.5% SDS for 10 min and re-probed with ³²P-labeled human β -actin control probe as an internal standard for the amounts of RNA loaded.

RESULTS AND DISCUSSION

Cloning of HAT cDNA—Using a pair of highly degenerate oligonucleotide primers, the partial 59-bp cDNA fragment for HAT, which contained a nucleotide sequence coding for the NH₂-terminal 19-residue amino acid sequence of the native HAT, was obtained by PCR amplification from human trachea cDNA. To stretch this cDNA sequence to the 3'-end, a 3'-RACE reaction was carried out. The resulting 0.9-kb amplified product was shown to encompass the entire nucleotide sequence of the 3' region, including the poly(A) tail of HAT cDNA (nucleotides 635–1517 in Fig. 1). The amino acid sequence deduced

² Y represents T or C; N represents C or I (inosine); R represents G or A; K represents G or T; M represents A or C; S represents G or C.

GAGTGGGAATCTCAAGCAGTTGAGTAGGCAGAAAAAGAACCTCTTCATTAAAGATTAA 60
 1 AATGTATAGGCCAGCAGCTGTAACTTGCAGTTCAAGATTTCAGAAATCAATGATGATG 120
 MYRPARRVVTSSTSR F L N P Y V V C
TTTCATTTGCTGCGCAGGGTAGTGATCTGGCAGTCACCATAGCTCTACTTGTCTACTT 180
 21 P I V V A G V V T L A V T T A L L V Y F
TTTAGCTTTGATCAAAAATCTTACTTTTATAGGAGCAGTTTCAACTCTCTAAATGTTGA 240
 41 L A P D Q K S Y F Y R S S F Q L L N V E
 ATATAATAGTCAGTTAAATTCACAGCTACACAGGAATACAGGACTTTGAGTGAAGAAAT 300
 61 Y N S Q L N S P A T Q E Y R T L S G R I
 TGAATCTCTGATTACTAAACATTCAAGAATCAAAATTTAAGAAATCAGTTTCATCAGAGC 360
 81 E S L I T K T F K E S N L R N Q F I R A
 TCAATGTTGCCAAACTGAGCAAGATGCTAGTGGTGTGAGAGCGGATGTTGTCTATGAAAT 420
 101 H V A K L R Q D G S G V R A D V V M K F
 TCAATTCACCTAGAAAATACAAATGGAGCATCAATGAAAAGCAGAAATGAGTCTGTTTACG 480
 121 Q F T R N N N G A S M K S I E S V L R
 ACAAATGCTGAATAACTCTGGAAACCTGGAAATAAACCTTCAACTGAGATAACATCACT 540
 141 Q M L N N S G N L E I N P S T B I T S L
 TACTGACCAAGCTGCAGCAAAATGCGTTTAAATGAATGTGGGCGGCTCAGACGTAAT 600
 161 T D Q A A A N W L I N E C G A G P D L I
 AACATTTCTGAGCAGAGAAATCTCTGGAGGCACTGAGGCTGAGGAGGAGGAGGCGCGTG 660
 181 T L S E Q R I L G G T E A R E G S W P W
 GCAAGTCAGCTGCGGCTCAATTAATGCCACCACTGTGGAGGCGGCTGATCAATACAT 720
 201 Q V S L R L N N A H H C G G S L I N N M
 GTGATCTCTGACAGCAGCTCACTCTTACAGCAACTCTAATCTCTGATGATGATGTC 780
 221 W I L T A A H C F R S N S N P R D W I A
 CACGTCTGGTATTTCCACAACATCTCTAACTAAGAATGAGAGTAAGAAATTTTAAAT 840
 241 T S G I S T T T F P K L R M R V R N I L I
 TCATAACAATTATAAATCTGCAACTCATGAAATGACATTGCACTTGTGAGACTTGAGAA 900
 261 H N N Y K S A T H E N D I A L V R L E N
 CAGTGTCACTTTACCAAGATATCCATAGTGTGTCTTCCAGCTGCTACCCAGAAATAT 960
 281 S V T F T K D I H S V C L P A A T Q N I
 TCCACCTGGCTCTACTGCTATGTAAACAGGATGGGGCGCTCAAGAAATATGCTGGCCAC 1020
 301 P P G S T A Y V T G W G A Q E Y A G H T
 AGTCCAGAGCTAAGGCAAGGACAGGTCAGAAATAAGTAATGATGTATGTAAATGCACC 1080
 321 V P E L R Q G Q V R I I S N D V C N A P
 ACATAGTTATAATGAGCCATCTTCTCTGGAATGCTGTGCTGCTGAGTACCTCAAGGTGG 1140
 341 H S Y N G A I L S G M L C A G V P Q G G
 AGTGGACCGATGTCAGGCTGCTGCTGGCCCACTAGTCAAGAAGACTACCGGCGCT 1200
 361 V D A C Q G D G S G G P L V Q E D S R L
 TTGGTTTATTTGGGATAGTAAAGCTGGGAGATCAGTGTGGCTGCGGGAATAGCCAGG 1260
 381 W F I V G I V S W G D Q C G L P D K P G
 AGTGTATCTCGAGTGACAGCTACCTTGACTGGAATAGGCAACAACTGGGATCTAGTG 1320
 401 V Y T R V T A Y L D W I R Q Q T G I *
 CAACAAGTGCATCCCTGTGTCAAAAGCTGTATGACAGGTGTGCTCTTAAATCCAAAG 1380
 CTTTACATTTCAACTGAAAAAGAACTAGAAATGTCTTAAATTTTACATTTTCTCC 1440
 ATATGGTTTAAACAACTGTTTAACTTTCTTTATTTAAAGGTTTCTATTTTCTCC 1500
 AAAAAAAAAAAAAA 1517

FIG. 1. Nucleotide sequence of HAT cDNA and its deduced amino acid sequence. The nucleotide sequence of the HAT cDNA is shown along with the deduced amino acid sequence beginning with the first ATG codon. A stop codon (TAG) at the terminus of the translation sequence is marked with an asterisk. Nucleotides are numbered at the right margin and amino acids on the left. The NH₂-terminal sequence obtained from the purified enzyme is underlined. The boxed amino acid sequence represents a potential transmembrane domain.

from this 0.9-kb fragment was shown to exactly contain the 15-amino acid sequence (amino acids 192–206 in Fig. 1) of the NH₂-terminal 20-amino acid sequence of the native HAT. With this 0.9-kb cDNA fragment as a probe, 1×10^6 clones of a human trachea λ gt10 cDNA library were then screened. Five of 28 independent positive clones were then subcloned and sequenced. The largest insert was shown to contain a 1323-bp sequence of cDNA (nucleotides 133–1455 in Fig. 1) but was considered not to contain the entire nucleotide sequence of the 5' region of HAT cDNA. To obtain the missing sequences in the 5' region of HAT cDNA, 5'-RACE reaction was carried out. The 5'-RACE procedure produced a 741-bp cDNA fragment (nucleotides 1–741 in Fig. 1). This product had a 609-bp nucleotide sequence overlapping (nucleotides 133–741 in Fig. 1) with the 5'-end of the largest insert of cDNA clone obtained by the cDNA library screening.

Sequence and Structural Features of HAT cDNA—Analysis of the cDNA clones obtained by the successive procedures in-

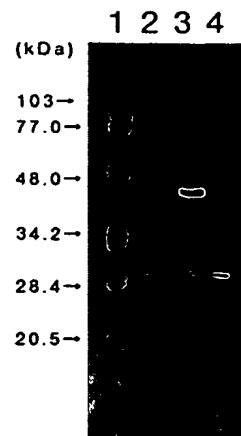


FIG. 2. Immunoblotting of the native HAT and the recombinant HAT. Specific binding was analyzed using the antibody against a peptide corresponding to the NH₂-terminal 19 amino acids of HAT as described previously (18). Lane 1, standard proteins; lane 2, purified native HAT (0.10 μ g); lane 3, lysate of infected Tn5 cells derived from a 20- μ l culture; and lane 4, purified recombinant HAT (0.10 μ g).

cluding 3'-RACE, cDNA library screening, and 5'-RACE showed a 1517-bp nucleotide sequence up to the poly(A) region (Fig. 1), which represented the HAT cDNA sequence. This nucleotide sequence was also shown to contain one open reading frame, and the polypeptide deduced from the cDNA included the 20-residue amino acid sequence of the NH₂ terminus of the native HAT (amino acids 187–206 in Fig. 1). The molecular mass of the polypeptide, including the NH₂ terminus of the 20 residues to the COOH terminus deduced from the stop codon TAG (nucleotide-1316), was estimated to be 25,308 Da. This value is similar to the apparent molecular mass (27 kDa) estimated by gel filtration of the native HAT protein purified from sputum (18).

In the 5'-flanking region of this cDNA, one in-frame stop codon TAG was located at nucleotide 26. Four in-frame ATG codons were detectable between this stop codon and the region encoding the native HAT, but none of these ATG codons satisfied the criteria for a Kozak consensus sequence (29). Therefore we could not determine the translational initiation site in the cDNA from the nucleotide sequence. To determine the initiation site, we expressed recombinant HAT in a baculovirus/insect cell system using the HAT cDNA. The recombinant virus containing the HAT cDNA was isolated, and the insect cell Tn5 was infected with the virus and then cultured. The lysate obtained by 1% Triton X-100 treatment of the infected cells was analyzed by immunoblotting with a rabbit antibody against a peptide corresponding to the NH₂-terminal 19-amino acid sequence of the native HAT (18) as primary antibody, and the immunoblotting indicated that the infected cells biosynthesized a protein with a molecular mass of 48 kDa as a main product (Fig. 2). The molecular mass of each polypeptide, deduced from the nucleotide sequence initiating from each of 4 ATG codons in the cDNA, was 46,263, 32,933, 31,436, and 30,107 Da, respectively. The molecular mass of 46,263 Da is the most similar to that of the recombinant protein expressed in the insect cells, suggesting that the ATG located nearest the 5'-end (at nucleotide 62) is the initiation codon of HAT.

To demonstrate that the cloned enzyme has the same activity as the native HAT, the recombinant HAT that was expressed in the baculovirus/insect cell system was isolated in its active form. The minor product in Fig. 2, lane 3 was isolated selectively as the active recombinant HAT from the infected cell lysate by sequential chromatographic procedures of the native

HAT purification (18). The purified recombinant enzyme has the molecular mass of 28 kDa on SDS-polyacrylamide gel electrophoresis and the identical 10 NH₂-terminal residues to the native HAT. Immunoblotting also showed the purified recombinant enzyme as same size as the native HAT (Fig. 2). The recombinant HAT had an enzymatic activity degrading fibrinogen, especially the α -chain (Fig. 3), similar to the native HAT. From these results, it was established that the isolated cDNA clone encodes HAT.

Based on these results, the nucleotide sequence of the cDNA for HAT (Fig. 1) was summarized as follows. The cDNA includes 1254 nucleotides coding for 418 amino acids and two untranslated nucleotide sequences composed of 61 and 185 nucleotides at the 5'-end and 3'-end, respectively. In the 3'-untranslated region, there is a polyadenylation signal sequence, ATTAAA, at nucleotides 1478–1483, 17 nucleotides distant from the poly(A) tail.

Analysis of Deduced Amino Acid Sequence of HAT—The open reading frame of HAT cDNA was thought to encode a polypeptide consisting of 418 amino acid residues, thus having the molecular mass of 46,263 Da. The NH₂-terminal 20-amino acid

sequence of the native HAT extends from Ile¹⁸⁷ to Leu²⁰⁶ in the sequence of the deduced polypeptide (Fig. 1). This result indicates that the Arg¹⁸⁶-Ile¹⁸⁷ peptide bond in the HAT polypeptide should be cleaved for activation of HAT. This type of cleavage has been shown to be a relatively common step for activation of many known serine protease zymogens (30, 31). Therefore it is likely that the HAT gene product is synthesized as a precursor protein that consists of a noncatalytic region with 186 amino acid residues (20,955 Da, amino acids 1–186 in Fig. 1) and a catalytic region with 232 amino acid residues (25,308 Da, amino acids 187–418 in Fig. 1) and that the precursor is converted to an active enzyme by limited proteolysis like trypsinogen to trypsin in the small intestine (32). In this noncatalytic region, there were two potential *N*-linked glycosylation sites, namely Asn-Asn-Ser and Asn-Pro-Ser, at Asn¹⁴⁴ and Asn¹⁵², respectively.

A hydropathy plot (22) of the predicted amino acid sequence of HAT precursor (Fig. 4) showed that a typical NH₂-terminal signal sequence (33–35) is not present, but a single obvious hydrophobic region (amino acids 13–43 in Fig. 1) is present near the NH₂ terminus. This hydrophobic region consisting of 31 amino acid residues does not contain any charged amino acids and is flanked by charged amino acids (Arg¹² and Asp⁴⁴). This internal hydrophobic region is thought to correspond to a transmembrane domain that anchors the protein to the cell membrane (36). A generalized rule in the eucaryotic transmembrane proteins (37, 38) suggests that the difference in total charge between 15-residue sequences on either side of the membrane-spanning hydrophobic region determines the orientation of the protein, with the more positive side facing the cytosol. As for the precursor polypeptide deduced from HAT cDNA, the NH₂-terminal side of the hydrophobic region had a net charge of +3, whereas the opposite side had that of +1. The charge on the NH₂-terminal side was +2, as positive as that on the COOH-terminal side. This result suggests that HAT precursor has an intracellular NH₂-terminal tail region consisting of 12 amino acid residues facing the cytosol and an extracellular COOH-terminal region consisting of 375 amino acid residues and containing the catalytic region. Therefore, the HAT precursor can be classified as a type II integral membrane protein (39, 40) and is thought to be synthesized as a membrane-bound precursor protein translocated to the cell surface, processed to a soluble form, and released.

Because neither the precursor nor intermediate form of HAT

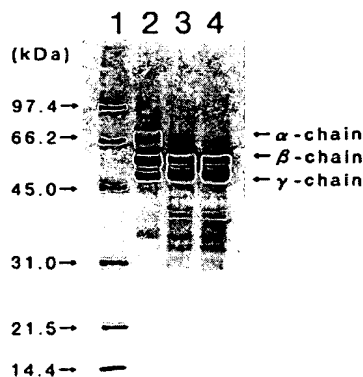


FIG. 3. Degradation of human fibrinogen by the native HAT and the recombinant HAT. Hydrolyzing reaction and SDS-polyacrylamide gel electrophoresis were done as described previously (18). For each reaction, 0.10 μ g of HAT was used. Lane 1, standard proteins; lane 2, fibrinogen (blank control); lane 3, fibrinogen hydrolyzed by native HAT; lane 4, fibrinogen hydrolyzed by recombinant HAT.

FIG. 4. Hydropathy plot of the deduced amino acid sequence of HAT. The method of Kyte and Doolittle (22) was used with averaging over a window of 10 residues. Hydrophobic residues show positive values, whereas hydrophilic residues show negative values. Amino acid numbering begins with the start codon Met.

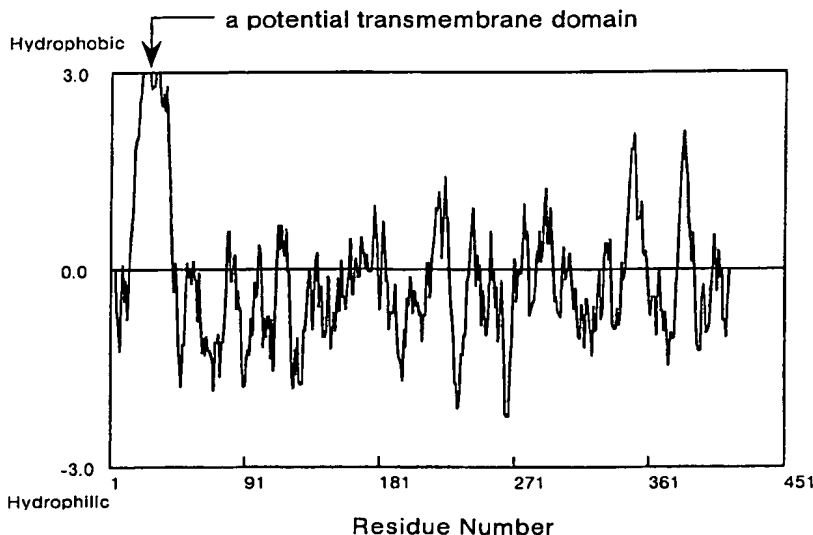


FIG. 5. Comparisons of the deduced amino acid sequence of the catalytic portion of HAT with those of other serine proteases. Identical amino acid residues are shaded, and the catalytic triad of histidine, aspartic acid, and serine are indicated by triangles. Hyphens represent gaps to bring the sequences to better alignment.

HAT	187: TLGGTEAEEGSHWVSI-----RLNNAHCGGSLNNMILTAAHCFRNSNP-RDW-I
Hepsin	: EVGGRTSLGRWHPVSI-----RYDGAHLGGSLSGDVLTAHCFPERNVLRSRWV
Enteropeptidase	: EVGGSNAKEGAPWVYGL--YYGGR--L-LEGASEVSSDHLVSAHGVYGRNLEPSKW-T
Acrosin	: EVGGKAAQHGAHPVMSLQIFRYNSHRYHTEGGSLNSRWVLTAAHG-FVGKNNVDMRL
Trypsin	: EVGGQEAAPSKHPVSI-----RVRDRYWHFEGGSLHPQVLTAAHCL-GPDV--KD--L
HAT	240: ATSGISTTFPK-LRMVRNLIHNNY---K-SATHE--NDIALVRLNSVTFKDIHSV
Hepsin	: FAGAVAQASPHGLQLGVQAVVYHGGYLPF-RDPNSEEKNDIALVHLSPLPLTEYIQPV
Enteropeptidase	: AILGLHMKSNLTSPQTPRLIDEIVINP---HYNRRRKNDIATMHLFKNYTDYIQPI
Acrosin	: VFGAKEITYGNKPKAPLQERYVEKIIIEKYN SATGNDIALVEITPPISCGRFIGPG
Trypsin	: ATLVRNSGTHLYYQQLLP-VSRIMVHP---QFYIIQTGADIALLEEPVNISSRVHTV
HAT	292: CLPAATQNI PPG-STAYVTGNG-AQEYAG-HTVPELRQGVRIISNDVC--N-APHSY--
Hepsin	: CLPAAGQALVDG-KICTVTGNG-NTQYYG-QQAGVLEARVPIISNDVC--N-GADFY--
Enteropeptidase	: CLPEENQVFPFG-RNCIAGNGTVVYQGT-TANI-LQEADVPLLSNERCQ-Q-QMPEY--
Acrosin	: CLPHFKAGLPRGSCQWVAGWYIEEKAP-RPSSILEARVDLIDLCLNS--TQWYNG
Trypsin	: MLPPASETFPPG-MPCWVTGNGVDNDDELPFPFLEKQVKVPIMENHICDAKYHLGAYTG
HAT	344: -NGAI-LSGMLCAQVPGGVDAQGGDGGPLV-QED-SR-RLWFIWGVSWGDOGLPDK
Hepsin	: -GNQI-KPKMFCAGYPEGGIDACQGGSGGPFVCEDSISRTPRWRLCGIVSWGTCALAQK
Enteropeptidase	: -N--IT-ENMICAGYEGGIDSCQGGSGGFLMC--QEN--NRWFLAGVTSFGYKCALPNR
Acrosin	: ---RVQPTNV-CAGYPVGKIDTCQGGSGGFLMC--KDSKESATVVVGGITSWGVCALAKR
Trypsin	: DDVRIIRDMLCAG--NSQRDSCKGGSGGFLVC--KVN--GTWLQAGVSWDEGCAQPNR
HAT	399: PGVYTRVTAYLDWI-RQQTGI-----
Hepsin	: PGVYTKVSDFREWI-FQAIKTHSEASGMVTQL-----
Enteropeptidase	: PGVYARVSRTFETI-QSFLH-----
Acrosin	: PGITYATWPLYLNIASKIGSNALRMISATPPPTTRPPPIRPPFSPHISAPLWYFQPP
Trypsin	: PGITYTRVTYLDWI-HHYVPKKP-----
Acrosin	: PRPLPPRPPAAQPPPPSPPPPPPPASPLPPPPPPPTPSSTTKLPQGLSFAKRLQQL
Acrosin	: IEVLKGKTYSDGKNHYDMETTELPELTSTS

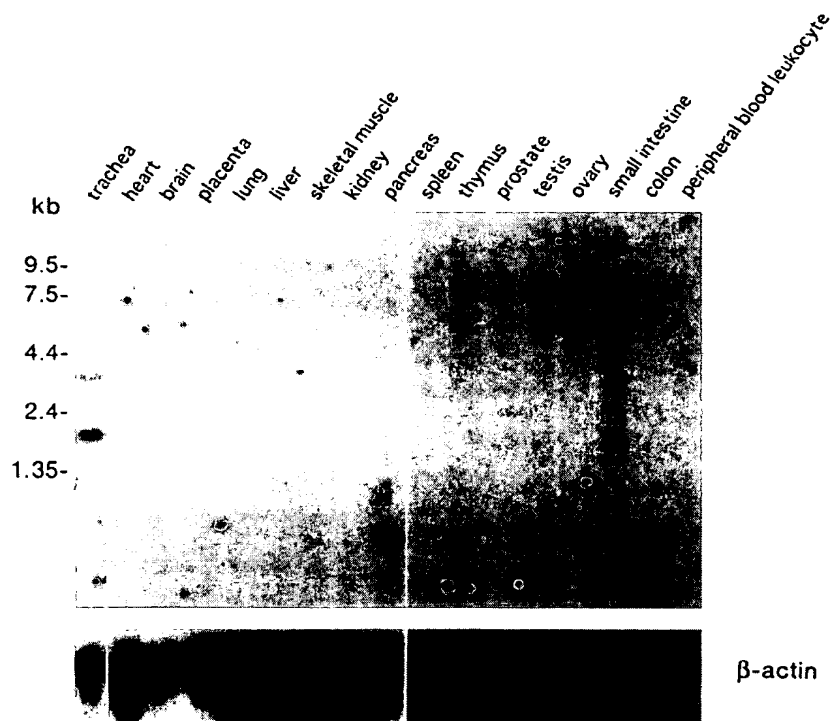


FIG. 6. Northern blot analysis of HAT mRNA in various human tissues. The blots were hybridized to HAT cDNA probe (upper panel). The same filters were re-hybridized with β -actin probe as an internal standard for the amounts of RNA loaded (lower panel).

has been isolated and characterized, it is unknown whether or not the membrane-bound HAT is active on the cell surface. The mechanisms of expression and activation of many serine proteases have been clarified. The predicted maturation process of HAT precursor described above is similar to that of the *Bacillus amyloliquefaciens* subtilisin (41). The subtilisin is synthesized as a membrane-associated precursor (preprosubtilisin) and released outside the cell after it is autocatalytically converted to an active form (42). Only mature subtilisin has been detected extracellularly (41). Active HAT contained in sputum samples was also detected extracellularly.

It is possible that the membrane-bound HAT or the portion

remaining in the membrane after release of the soluble HAT may be involved in some important physiological processes on the cell surface through interaction with ligands, other proteins, or the surface. Recent reports have shown that some viruses and a bacterial toxin utilize cell surface proteases as receptors (43–47), indicating other usage in addition to intrinsic roles of these proteins.

Homology of Amino Acid Sequence of HAT with Other Proteases—To find any similarity in the primary structure between HAT and known proteins, we surveyed publicly available data banks. Previous investigators have shown that the serine protease family has a common catalytic site consisting of

three amino acid residues, His, Asp, and Ser, joined by hydrogen bonds to display catalytic action as a catalytic triad, although they are located apart from each other in the primary structure of the enzyme (48). Based on these established facts, the catalytic site of HAT is thought to consist of amino acid residues His²²⁷, Asp²⁷², and Ser³⁶⁸ (Fig. 5). In comparison of the amino acid sequence of HAT with those of other serine proteases, the most striking similarity was found around this putative catalytic triad as shown in Fig. 5. Six of seven cysteine residues in the catalytic region of HAT were at identical positions as those of other serine proteases (Fig. 5). Nine cysteine residues were contained in the deduced polypeptide of HAT precursor, and the Cys²⁰ was located in the predicted transmembrane domain. Based on the locations of the known disulfide bridges in other serine proteases (49), it is postulated that the other eight cysteine residues may form four disulfide bonds, which are located at cysteine pairs 212/228, 337/353, and 364/393 in the catalytic region and at 173/292 between the noncatalytic region and the catalytic region.

It was shown that the amino acid sequence of the catalytic region of HAT was homologous to that of the other human serine proteases: 38% identity with hepsin (50), 32% with enteropeptidase (51), 30% with acrosin (52), and 29% with mast cell tryptase (53). Hepsin, of which the catalytic region shows the highest similarity with that of HAT in this survey, is a cell surface protease widely expressed in various tissues including liver and is suggested to play a role in cell growth and maintenance of cell morphology (54).

On the other hand, the amino acid sequence of the noncatalytic region of HAT showed no significant similarity with those of other proteins and had neither kringle nor an EGF-like domain, which are found in some kinds of proteases relating to the blood coagulation, fibrinolysis, and complement cascades (55). The function or roles of this unique and relatively long noncatalytic portion of HAT precursor are unknown.

Northern Blot Analysis—Previously, we showed immunohistochemically that HAT protein was expressed in the cells of submucosal serous glands of human bronchi and trachea (18). Serous glands are widely distributed in various human tissues. Therefore multiple tissue Northern blot analysis was carried out to confirm that HAT mRNA was expressed in the human lower airway and also to clarify whether or not HAT mRNA was expressed in human other tissues. As shown in Fig. 6, a 1.9-kb transcript was detectable in only the trachea blot among the 17 different types of tissues examined, such as heart, brain, pancreas, lung, and liver. The mRNA size is in fairly good accordance with that (1517 bp) of the HAT cDNA established in the present work. In addition to the 1.9-kb mRNA, 3.0-kb and 0.9-kb signals were weakly detectable in the trachea and pancreas blot, respectively. These two transcripts may appear as result of an alternative splicing/polyadenylation process or represent a cross-hybridizing mRNA, but the nature of these transcripts is unknown. These results strongly suggest that HAT mRNA is more actively expressed in the lower airway including trachea than in the other tissues examined and support our previous result that HAT is localized in cells of submucosal serous glands of trachea and bronchi.

Although the native HAT was found in the sputum of patients with chronic airway diseases, HAT mRNA is thought to be expressed in the normal tissues of healthy subjects, because the trachea poly(A)⁺ RNA subjected to the Northern blot was obtained from the normal trachea tissues of three white male subjects who died of trauma or acute heart failure. It will be useful to compare expression levels of mRNA and protein of HAT in the patients with airway diseases with those in healthy subjects to clarify the physiological and pathophysiological sig-

nificance of HAT in the airway. In the airway, various kinds of proteins such as lysozyme (56), secretory IgA (57), and secretory leukocyte protease inhibitor (58) are secreted from the submucosal serous glands onto mucous membrane and become constituents of airway mucous or bronchial secretions (59). These proteins play important roles in the host defense system of airways together with respiratory mucous glycoproteins, which are secreted from mucous glands cells and goblet cells (59). HAT may be released from the serous glands with these proteins and play some biological role in the host defense system on the mucous membrane independently of or in cooperation with other substances in airway mucous or bronchial secretions.

In summary, it was confirmed through the present work that HAT is a novel trypsin-like serine protease by analyzing the primary structure of the polypeptide deduced from the nucleotide sequence of its cDNA. However, the mechanism of activation of the HAT precursor to mature enzyme, the physiological role of the enzyme, and biological significance of the noncatalytic region of the precursor remain to be resolved.

REFERENCES

- Cohen, A. B. (1983) *Am. Rev. Respir. Dis.* **127**, (suppl.) 2–58
- Janoff, A. (1985) *Am. Rev. Respir. Dis.* **132**, 417–433
- Suter, S., and Chevallier, I. (1991) *Eur. Respir. J.* **4**, 40–49
- O'Connor, C. M., Gaffney, K., Keane, J., Southey, A., Byrne, N., O'Mahoney, S., and Fitzgerald, M. X. (1993) *Am. Rev. Respir. Dis.* **148**, 1665–1670
- Meyer, K. C., Lewandoski, J. R., Zimmerman, J. J., Nunley, D., Calhoun, W. J., and Dopic, G. A. (1991) *Am. Rev. Respir. Dis.* **144**, 580–585
- Lee, C. T., Fein, A. M., Lippmann, M., Holtzman, H., Kimbel, P., and Weinbaum, G. (1981) *N. Engl. J. Med.* **304**, 192–196
- Schwartz, L. B., Irani, A. M., Roller, K., Castells, M. C., and Schechter, N. M. (1987) *J. Immunol.* **138**, 2611–2615
- Caughey, G. H. (1994) *Am. J. Respir. Crit. Care Med.* **150**, (suppl.) 138–142
- Gervasoni, J. E., Jr., Conrad, D. H., Hugli, T. E., Schwartz, L. B., and Ruddy, S. (1986) *J. Immunol.* **136**, 285–292
- Pejler, G., and Karlström, A. (1993) *J. Biol. Chem.* **268**, 11817–11822
- Saarienen, J., Kalkkinen, N., Welgus, H. G., and Kovanen, P. T. (1994) *J. Biol. Chem.* **269**, 18134–18140
- Nakano, A., Kishi, F., Minami, K., Wakabayashi, H., Nakaya, Y., and Kido, H. (1997) *J. Immunol.* **159**, 1987–1992
- Janoff, A. (1983) *J. Appl. Physiol.* **55**, 285–293
- Sibille, Y., and Reynolds, H. Y. (1990) *Am. Rev. Respir. Dis.* **141**, 471–501
- Kido, H., Yokogoshi, Y., Sakai, K., Tashiro, M., Kishino, Y., Fukutomi, A., and Katunuma, N. (1992) *J. Biol. Chem.* **267**, 13573–13579
- Sakai, K., Kawaguchi, Y., Kishino, Y., and Kido, H. (1993) *J. Histochem. Cytochem.* **41**, 89–93
- Tashiro, M., Yokogoshi, Y., Tobita, K., Seto, J. T., Rott, R., and Kido, H. (1992) *J. Virol.* **66**, 7211–7216
- Yasuoka, S., Ohnishi, T., Kawano, S., Tsuchihashi, S., Ogawara, M., Masuda, K., Yamaoka, K., Takahashi, M., and Sano, T. (1997) *Am. J. Respir. Cell Mol. Biol.* **16**, 300–308
- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
- Hanahan, D. (1983) *J. Mol. Biol.* **166**, 557–580
- Birnboim, H. C., and Doly, J. (1979) *Nucleic Acids Res.* **7**, 1513–1523
- Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132
- Smith, T. F., and Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197
- Frohman, M. A., Dush, M. K., and Martin, G. R. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85**, 8998–9002
- Feinberg, A. P., and Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13
- Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K. (1997) *Current Protocols in Molecular Biology*, Suppl. **18**, 16.9.1–16.11.12, John Wiley & Sons, Inc., New York
- Wickham, T. J., Davis, T., Granados, R. R., Shuler, M. L., and Wood, H. A. (1992) *Biotechnol. Prog.* **8**, 391–396
- Lehrack, H., Diamond, D., Wozney, J. M., and Boedtker, H. (1977) *Biochemistry* **16**, 4743–4751
- Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8148
- Blow, D. M. (1971) in *The Enzymes* (Boyer, P. D., ed) 3rd Ed., Vol. 3, pp. 185–212, Academic Press, Inc., New York
- Keil, B. (1971) in *The Enzymes* (Boyer, P. D., ed) 3rd Ed., Vol. 3, pp. 250–275, Academic Press, Inc., New York
- Davie, E. W., and Neurath, H. (1955) *J. Biol. Chem.* **212**, 515–529
- von Heijne, G. (1983) *Eur. J. Biochem.* **133**, 17–21
- Wickner, W. (1980) *Science* **210**, 861–868
- von Heijne, G. (1985) *J. Mol. Biol.* **184**, 99–105
- von Heijne, G., and Manoil, C. (1990) *Protein Eng.* **4**, 109–112
- Hartmann, E., Rapoport, T. A., and Lodish, H. F. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 5786–5790
- Parks, G. D., and Lamb, R. A. (1991) *Cell* **64**, 777–787
- High, S. (1992) *Bioessays* **14**, 535–540
- Semenza, G. (1986) *Annu. Rev. Cell Biol.* **2**, 255–313
- Wells, J. A., Ferrari, E., Henner, D. J., Estell, D. A., and Chen, E. Y. (1983)

- Nucleic Acids Res.* 11, 7911-7925
42. Power, S. D., Adams, R. M., and Wells, J. A. (1986) *Proc. Natl. Acad. Sci. U. S. A.* 83, 3096-3100
43. Kido, H., Fukutomi, A., and Katunuma, N. (1991) *Biomed. Biochim. Acta* 50, 781-789
44. Delmas, B., Gelfi, J., L'Haridon, R., Vogel, L. K., Sjöström, H., Noten, O., and Laude, H. (1992) *Nature* 357, 417-420
45. Yeager, C. L., Ashmum, R. A., Williams, R. K., Cardellicchio, C. B., Shapiro, L. H., Look, A. T., and Holmes, K. V. (1992) *Nature* 357, 420-422
46. Söderberg, C., Giugni, T. D., Zaia, J. A., Larsson, S., Wahlberg, J. M., and Möller, E. (1993) *J. Virol.* 67, 6576-6585
47. Knight, P. J. K., Crickmore, N., and Ellar, D. J. (1994) *Mol. Microbiol.* 11, 429-436
48. Kraut, J. (1977) *Annu. Rev. Biochem.* 46, 331-358
49. Young, C. L., Barker, W. C., Tomaselli, C. M., and Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) Vol. 5, pp. 73-93, National Biochemical Research Foundation, Washington, D. C.
50. Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* 27, 1067-1074
51. Kitamoto, Y., Veile, R. A., Donis-Keller, H., and Sadler, J. E. (1995) *Biochemistry* 34, 4562-4568
52. Adham, I. M., Klemm, U., Maier, W. M., and Engel, W. (1990) *Hum. Genet.* 84, 125-128
53. Miller, J. S., Westin, E. H., and Schwartz, L. B. (1989) *J. Clin. Invest.* 84, 1188-1195
54. Kurachi, K., Torres-Rosado, A., and Tsuji, A. (1994) *Methods Enzymol.* 244, 100-114
55. Patthy, L. (1990) *Blood Coagul. Fibrinolysis* 1, 153-166
56. Bowes, D., and Corrin, B. (1977) *Thorax* 32, 163-170
57. Goodman, M. R., Link, D. W., Brown, W. R., and Nakane, P. K. (1981) *Am. Rev. Respir. Dis.* 123, 115-119
58. de Water, R., Willems, L. A., van Muijen, G. N., Franken, C., Fransen, J. A., Dijkman, J. H., and Kramps, J. A. (1986) *Am. Rev. Respir. Dis.* 133, 882-890
59. Kaliner, M., Shelhamer, J. H., Borson, B., Nadel, J., Patow, C., and Marom, Z. (1986) *Am. Rev. Respir. Dis.* 134, 612-621



Exhibit 44

Corin, a Mosaic Transmembrane Serine Protease Encoded by a Novel cDNA from Human Heart*

(Received for publication, January 11, 1999)

Wei Yan, Ning Sheng, Marian Seto, John Morser, and Qingyu Wu†

From the Departments of Cardiovascular Research and Biophysics, Berlex Biosciences, Richmond, California 94804

A novel cDNA has been identified from human heart that encodes an unusual mosaic serine protease, designated corin. Corin has a predicted structure of a type II transmembrane protein and contains two frizzled-like cysteine-rich motifs, seven low density lipoprotein receptor repeats, a macrophage scavenger receptor-like domain, and a trypsin-like protease domain in the extracellular region. Northern analysis showed that corin mRNA was highly expressed in the human heart. In mice, corin mRNA was detected by *in situ* hybridization in the cardiac myocytes of the embryonic heart as early as embryonic day (E) 9.5. By E11.5–13.5, corin mRNA was most abundant in the primary atrial septum and the trabecular ventricular compartment. Expression in the heart was maintained through the adult. In addition, mouse corin mRNA was also detected in the prehypertrophic chondrocytes in developing bones. By fluorescent *in situ* hybridization analysis, the human corin gene was mapped to 4p12–13 where a congenital heart disease locus, total anomalous pulmonary venous return, had been previously localized. The unique domain structure and specific embryonic expression pattern suggest that corin may have a function in cell differentiation during development. The chromosomal localization of the human corin gene makes it an attractive candidate gene for total anomalous pulmonary venous return.

Serine proteases are essential for a variety of biological processes including food digestion, complement activation, and blood coagulation (1–3). In *Drosophila*, serine proteases are also involved in developmental pathways. For example, serine proteases encoded by the *nudel*, *gastrulation defective*, *easter*, and *snake* genes are key components of a proteolytic cascade that is critical for the establishment of the dorsal-ventral pattern in developing embryos (4–6). Genetic defects in these genes often lead to the disruption of the dorsal-ventral axis, resulting in embryonic lethality (7).

Most serine proteases of the trypsin family are secreted proteins. Several members from this family have been identified that contain an integral transmembrane domain. Hepsin, for example, is a serine protease expressed on the surface of hepatocytes. Structurally, hepsin is a type II transmembrane protein with the transmembrane domain at its amino terminus and the protease domain at the carboxyl terminus exposed to

the outside of the cell (8). In tissue culture studies, hepsin was shown to contribute to hepatocyte growth (9). However, the physiological significance of the growth stimulating activity of hepsin remains unknown (10). In *Drosophila*, Stubble-stubloid protein, another transmembrane serine protease, shares structural similarities with hepsin (11). Genetic studies demonstrated that Stubble-stubloid is essential for epithelial morphogenesis and development of the fruit fly. Defects in the *Stubble-stubloid* gene cause malformation of legs, wings, and bristles. Most recently, other transmembrane serine proteases were isolated and cloned from human trachea and small intestine (12, 13). The biological function of these newly discovered membrane-bound serine proteases has not yet been determined.

In this study, we report the cloning of a cDNA from the human heart that encodes a novel transmembrane serine protease, designated corin. Corin has a predicted structure of a type II transmembrane protein containing two frizzled-like cysteine-rich motifs, seven LDL¹ receptor repeats, a macrophage scavenger receptor-like domain, and a trypsin-like protease domain in the extracellular region. *In situ* hybridization revealed that corin mRNA was expressed in the embryonic heart as early as E9.5, and the expression in the heart was maintained through the adult stage. In addition, corin mRNA was detected in prehypertrophic chondrocytes of the developing bones. The unusual domain structures and specific expression pattern suggested that corin may have a function in cell differentiation during embryonic development.

EXPERIMENTAL PROCEDURES

Materials—Human cancer cell lines, HEC-1-A (endometrium adenocarcinoma), U2-OS (osteosarcoma), SK-LMS-1 (vulva sarcoma), RL95-2 (endometrium carcinoma), and AN3-CA (endometrium adenocarcinoma) were obtained from the American Type Culture Collection (ATCC). Human heart cDNA libraries and human and mouse multiple tissue Northern blots were purchased from CLONTECH (Palo Alto, CA). Mouse tissue sections used for *in situ* hybridization were purchased from Novagen (Madison, WI). Tissue culture media and supplements were from Life Technologies Inc. All other chemicals were obtained from Sigma.

Isolation of Human Corin cDNA Clones—An expressed sequence tag (EST) clone was found in a human heart cDNA library from the Incyte EST data base that shared significant sequence homology with trypsin, indicating that the EST may encode a novel serine protease gene. A 2.1-kb *EcoRI-XhoI* insert from this EST clone was used to screen a human heart cDNA library (CLONTECH). Approximately, 5×10^6 lambda phage clones were screened, and two positive clones were isolated that contained inserts of 3.5 and 3.1 kb, respectively. The DNA sequences of these two clones were determined. Oligonucleotide prim-

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AF133845.

† To whom correspondence should be addressed: Berlex Biosciences, 15049 San Pablo Ave., Richmond, CA 94804. Tel.: 510-669-4737; Fax: 510-669-4246; E-mail: qingyu_wu@berlex.com.

¹ The abbreviations used are: LDL, low density lipoprotein; EST, expressed sequence tag; FISH, fluorescent *in situ* hybridization; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; ORF, open reading frame; RT, reverse transcriptase; PCR, polymerase chain reaction; RACE, rapid amplification of cDNA ends; TAPVR, total anomalous pulmonary venous return; kb, kilobase pair; bp, base pair; E, embryonic day.

ers were designed to clone further 5' end cDNA sequences by 5' rapid amplification of cDNA ends (RACE) using Marathon-ready human heart cDNA templates (CLONTECH). The PCR products from 5' RACE were cloned into pCRII vector (Invitrogen, San Diego, CA) and sequenced. Oligonucleotide primers used in the 5' RACE experiments were 5'-CAGTTGGTTTGAACAAGTGCAGGG-3', 5'-TGCAAGGAGG-GATACGCTCGCCTG-3', 5'-AATCCCAAGAACAGACTCACAGCG-3', 5'-CGGGTCACAGAGAGAGTACCAC-3', 5'-GGTCTCCTTCTTGA-CATGAATCTG-3', 5'-CGGAGCCCCATGAAGTTAAACCA-3', and 5'-AACAAAAGGATCCTTGGAGGTCGGACGAGT-3'. The final 5' end sequence of human corin cDNA was derived from at least three independent clones. The full-length cDNA sequence was compiled using the Genetics Computer Group (GCG) software (version 9.1, Madison, WI).

Northern Analysis—Northern blots containing poly(A)⁺ RNA samples (2 µg/lane) from multiple human and mouse tissues were purchased from CLONTECH. Human and mouse corin cDNA probes were labeled with [³²P]dCTP using a random primed DNA labeling kit (Roche Molecular Biochemicals). Northern hybridization was performed at 42 °C overnight in a solution containing 40% formamide, 5× Denhardt's solution, 6× SSC, 100 µg/ml salmon sperm DNA, and 0.1% SDS. Blots were washed with 0.2× SSC, 0.1% SDS at 60 °C and then exposed to Fuji imaging plates. As a control, the blots were reprobed with a human actin cDNA probe provided by CLONTECH.

RT-PCR—mRNA samples were isolated from Hec-1-A, U2-OS, SK-LMS-1, and AN3-CA cells using a commercial RNA preparation kit (Oligotex Direct mRNA Mini Kits, Qiagen). First strand cDNAs were synthesized using SuperScript II RNase[−] reverse transcriptase (Life Technologies Inc.). Human corin-specific oligonucleotide primers (sense primer, 5'-AACAAAAGGATCCTTGGAGGTCGGACGAGT-3', and antisense primer, 5'-CGGAGCCCCATGA AGTTAATCCA-3') were used to amplify a 630-bp fragment of corin cDNA between nucleotides 2475 and 3105. Oligonucleotide primers TFR1 (5'-GTCAATGTCCCAACCGT-CACCAGA-3') and TFR2 (5'-ATTTCGGGAATGCTGAGAAAAACAGACAGA-3'), derived from the human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene, were used as an internal quantification control. PCR reactions were performed with a thermal cycler (Perkin-Elmer, model 480). PCR products were separated on 1% agarose gels and visualized by ethidium bromide staining.

In Situ Hybridization—Mouse adult heart and embryonic tissue sections were deparaffinized in xylene, rehydrated, and fixed in 4% paraformaldehyde. The tissues were digested with proteinase K (20 µg/ml), then treated with triethanolamine/acetanhydride, and dehydrated. An 800-bp mouse corin cDNA fragment from the coding region was cloned into pCRII (Invitrogen) in two orientations to yield plasmids pM11 and pM41. The plasmids were linearized by *Hind*III digestion. Sense and antisense probes were synthesized using T7 RNA polymerase (T7/SP6 transcription kit, Roche Molecular Biochemicals) and labeled with [³³P]UTP (Amersham Pharmacia Biotech). The hybridization was carried out as described (14). The slides were dehydrated and dipped in Kodak NTB-2 emulsion and exposed for 4 weeks in light-tight boxes at 4 °C. Photographic development was carried out in a Kodak D-19 developer. The slides were stained with hematoxylin/eosin and analyzed using both light- and dark-field optics of a Zeiss microscope.

Fluorescent in Situ Hybridization (FISH) Analysis—P1 phage clones containing the human corin gene were isolated by filter hybridization using a human corin cDNA as the probe. One clone was confirmed by DNA sequencing using a primer from human corin cDNA. The DNA fragment from this P1 phage was labeled with digoxigenin-dUTP. The labeled probe was combined with sheared human DNA and hybridized to metaphase chromosomes derived from PHA-stimulated peripheral blood lymphocytes in a solution containing 50% formamide, 10% dextran sulfate, and 2× SSC. Hybridization signals were detected by fluorescent-labeled antidigoxigenin antibodies and counter-staining with 4,6-diaminodino-2-phenylindole. A total of 80 metaphase cells were analyzed of which 74 cells exhibited specific labeling.

Homology Model of the Protease Domain of Corin—A model of the corin protease domain (amino acids 802–1042) was built based on the structure of bovine chymotrypsinogen A at 1.8-Å resolution (15, 16), using the homology program (Insight II, 1995, MSI, San Diego, CA). Rotamers were used for non-identical side chain replacements (16). Coordinates for the loop insertions were extracted from the Brookhaven protein data bank (17). The model was refined by energy minimization using the AMBER force field (Discover 95.0), with a distance-dependent dielectric constant. The minimization used the steepest descents and conjugate gradient methods as follows: first for the loops only where insertions and deletions occurred, then side chains, and a final round of minimization keeping the Cα atoms fixed. The residues of corin (His⁸⁴³,

Asp⁸⁹², and Ser⁹⁸⁵) corresponding to the catalytic triad of the template structure were also held fixed.

RESULTS

Cloning of the Full-Length Human Corin cDNA—A computer search using the BLAST program identified an EST clone from a human heart library that shared significant homology with serine protease family members, such as trypsin. The EST clone was used to isolate the full-length cDNA of a novel gene, designated corin for its abundant expression in the heart. The sequence of the full-length corin cDNA, 4933 bp in length, is shown in Fig. 1. The size of the cDNA is consistent with the length of corin mRNA (~5 kb) detected by Northern analysis (Fig. 4A). An ATG codon is located at position 95 that may represent the translation initiation site. The open reading frame (ORF) spans 3126 bp with a 5'-untranslated region of 94 nucleotides before the initiation codon. At the 3' end, there is a 1.7-kb 3'-untranslated region after the stop codon at position 3221. A polyadenylation signal of AATAAA is present 12 nucleotides before the poly(A)⁺ tail.

The Domain Structure of Human Corin—The ORF of the human corin cDNA encodes a polypeptide of 1042 amino acids with a calculated mass of 116 kDa. At the amino terminus of the predicted corin protein, there is no discernible signal peptide sequence. Hydropathy plots using the GCG program identified a highly hydrophobic region between amino acids 46 and 66 (Fig. 2B). This hydrophobic sequence could serve as a potential transmembrane domain. There are positively charged amino acid residues immediately preceding the putative transmembrane segment, suggesting that corin is a type II transmembrane protein with the amino terminus present in the cytosol (18). Consistent with this hypothesis, there are 19 predicted N-linked glycosylation sites present in the extracellular domains of corin (Fig. 1).

Analysis of the corin protein sequence showed that in the extracellular region there are two frizzled-like cysteine-rich domains, seven LDL receptor repeats, one macrophage scavenger receptor-like domain, and one trypsin-like serine protease domain (Fig. 2A). As shown in Fig. 2A, two frizzled-like cysteine-rich domains are located at amino acids 134–259 and 450–573, respectively. Amino acid sequences of these two domains share significant similarities with the extracellular cysteine-rich domain of the *Drosophila* Frizzled protein, a seven-transmembrane receptor essential for polarity determination during the development of the fruit fly (19). The frizzled-like cysteine-rich domains have also been found in other proteins, such as Dfz2 in *Drosophila* (20), Lin-17 in *Caenorhabditis elegans* (21), and FZ-1 in human (22). The sequences of the two frizzled-like cysteine-rich domains in corin are closest to those in Lin-17 and FZ-1. As shown in Fig. 2C, all the 10 conserved cysteine residues are present in the frizzled-like cysteine-rich domains of corin.

Between amino acids 268–415 and 579–690 (Fig. 2, A and D), there are seven cysteine-rich repeats homologous to the LDL receptor class A repeats (23). Each repeat is about 36 amino acids long and contains six cysteine residues as well as a highly conserved cluster of negatively charged amino acids. In the LDL receptor, these cysteine-rich repeats bind calcium ions and play an essential role in endocytosis of the extracellular ligands (23). Similar motifs have been found in the extracellular domain of other membrane receptors, such as LDL receptor-related protein (LRP1) (24), megalin (also known as LRP2 or gp330) (25), complement proteins (26), enterokinase (27), and *Drosophila* proteins yolkless and nudel (28, 29).

In addition to the frizzled-like cysteine-rich domains and LDL receptor-like repeats, there is another cysteine-rich region between amino acids 713 and 801 in corin (Fig. 2, A and E).

1	AAATCATCCGAGTGCCTCCCGGGGACACCTAGAGCAGCAAAAGGACCAAGATAAA	60	2281	ATGGCTTTAGGAGAACATCTGTGACCAAAATGATACAGGACAGGACAAAGAGCGCGG	2340
61	AGTGGACAGAAATAAGCGAGACTTTTATCCATCAACAGCTCTCGCCCTGCTCCG	120	730	H G L G E P S V T K L I Q E Q E K E P R	749
	<u>H K Q S P A L A P</u>	9			
121	GAAGAGCGCTACCGCAGAGCGGGTCCCAAGCGGCTCTGAGAGCTGATGACAAATAC	180	2341	TGGCTGACATTACACTCCAACTGGGAGCGCTCAATGGGACCACTTTACATGAATCTTA	2400
10	B E R Y R R R A G S P K P V L R A D D M N	29	750	W L T L E S N W E S L H G T Y L H E L L	769
181	ATGGGAGGCTGCTCTCAGAAGCTGGCGACTGCTAACTCTCCGGTCTCTATTGCTG	240	2401	GTAATGGGAGCTCTGTGAGAGCAGAAATTTCTCTCTGCTGATCAAAACAGAC	2460
30	M G N G C S Q K L A T A N L L R L L L L	49	770	V N G G Q S C E S R S K I S L L C T K Q D	789
241	GTCCTGATTCACTGATCTGCTCTGCTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG	300	2461	TGTGGGCGCCCGCTGCTGCCGAATGAACAAAGGATCTTGAGAGTGGGAGGAGTGG	2520
50	<u>Y L I P C Y C A L V L L L L Y L L L</u> S Y V	69	790	C G R R P A A R M H K R I L G G R T S R	809
301	GGAAACATTACAAAGGCTCTATTAAATCAATGGGAGTGAACCTTGGTCACTGATGGT	360	2521	CTGGAGGTGGGCACTGGCAGTGTCTCTGAGAGTGAACCCAGTGGACATATCTGTGGC	2580
70	G T L O K V Y F K S H G S E P L V T D G	89	810	P G R W P W Q C G L Q S E P S G H I C G	829
361	GAATCCAAAGGTCGATGTTATCTTCAAAATCAAAATTTAACCAGGACACTGCTGGTG	420	2581	TGTGCTCTATTGCCAAGATGGGTCTGACAGTTCGCCCACTGCTCGAGGGGAGAGAG	2640
90	E I O G S D V I L T N T I Y N Q S T V V	109	830	C V L I A K K N V L T V A H C F E G R E	849
421	TCTATGTCACATCCGACCAAGCTTCCAGCTGGACTACGGATGCTCTCTCCCAAGG	480	2641	AATGTCGACTTTGGAAGTGGTCTGGCATCAACATCTAGACCATCATCAGTGTTC	2700
110	S T A H P D O H V P A W T T D A S L P G	129	850	N A A Y W K V V L G I N H L D H P S V F	869
481	GACCAAGTCACAGGAATACAGTGCCTGTATGAACATCAACCAAGCAGTGTGATG	540	2701	ATGCACACGCTTTGTGAAGACCATCATCTGCATCCCGCTGACAGTGGAGCTGGTG	2760
130	D O S H R H T S A C M H I T H S C C O M	149	870	N O T R F V K T I I L R P R Y S R A V V	889
541	CTGCCCTACCAAGCCAGCTGACACCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	600	2761	GACTATGACATCAGCATCTGAGCTGAGTGAAGACATCAGTGAAGTGGTACGTCGG	2820
150	L P Y H A T L T P L L S V Y N H E A E	169	890	D Y D I S I V E L S E D I S E T G Y V R	909
601	AAGTCTCTCAAGTTTTCACATCTCCATGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	660	2821	CTGTCTGCTGCTGCCCAACCCGAGCAGTGGCTAGAGCTGCACCTGCTCTATATCACA	2880
170	K F L K F T Y L H R L S C Y Q H I M L	189	910	P V C L P H P E Q W L E P D T Y C Y I T	929
661	TTTGGCTGACCTCGCT	720	2881	GGCTGGGCGCACATGGGCAATAAAATGCCATTAAAGTGCAGAGCGGCTGGCCTCT	2940
190	F G C T L A F P E C I I D G D D S H G L	209	930	G W G H M G N K M P F K L Q E G E V R I	949
721	CTGCCCTGATGCT	780	2941	ATTTCTCTGGAACATTGTCAGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	3000
210	L P C R S P F C E A A K E G C E S V L G M	229	950	I S L E H C Q S Y F D M K T I T T R M I	969
781	GTAATTAATCTCTGGCGGATTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	840	3001	TGTGCTGGCTATGACTCTGGCAGCTTGATTCATGCTGCTGCTGCTGCTGCTGCTGCT	3060
230	V H Y S W P D F L R C S Q F R N Q T E S	249	970	C A G Y E S G T V D S C H G D S G G P L	989
841	AGCAATGTCAGCAGAATTTGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	900	3061	GTTTGTGAGAAGCTGGGAGCGCTGGACATTAATTTGGAATTAATTTATGAGGCTCGTC	3120
250	S H V Y S R I C F S P Q Q E N G K O L L C	269	990	V C E K P G G R W T L F L G S W G G S V	1009
901	GGAAAGGTCAGAACTTTCT	960	3121	TGCTTTTCCAAAGTCTGGGCGCTGGGCTTTATAGTAATGTGCATATTTCTGCTGAATGG	3180
270	G R G E N F L C A S G I C I P G K L O C	289	1010	C F S K V L G P G V Y S H V S Y F V E W	1029
961	AATGGCTACCAAGCTGTGACGAGTGAAGTGAAGGCTCATTTGCACTGACGAGCAGAT	1020	3181	ATTAAAGACAGATTTACATCCAGACTTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	3240
290	N G Y N D C T G D D W S D E A H C N C S E N	309	1030	I K R Q I Y I Q T F L L N	
1021	CTGTTCATCTCTACACAGGCAAGTCCCTTAATACAGCTCTGTGTGTGTGTGTGTGTGT	1080	3241	ACTTTTGGCAGCTACACTAAAGAAATGGCTCTCTGACTGTGAAGAGTGGCTGCAGA	3300
310	L F H C R T G K C L H Y S L V C D G Y D	329	3301	GAGCTGTACAGAGCACTTTTCAATGGCAGAAATGCTCAATCTGCACTGCAAAATTTGCA	3360
1081	GACTGTGGGATTTGAGTGTGAGCAGAACTCTGATTGCAATCCCAACAGAGCATGCG	1140	3361	TGTTGTTTGGACTAATTTTTCATTTATTTTTCACCTCAATTTTCTCTCTCTCTCTCTCT	3420
330	D C G D L S D E Q N C D C N P T T E R R	349	3421	AACTTCAATGAAGACTTTTCAAAAGCAAAAGAGAGCTTTGCTCTTTTGGCAGGCTCT	3480
1141	TGCGGGACCGGCGCTGCATCCGATGGAGTGGTGTGTGTGTGTGTGTGTGTGTGTGTGT	1200	3481	AACTATGCTGAGCAGCAAAATATGCACTCTGGCGAGTTTAAATCAGGTGCTACAGT	3540
350	C G D G R C I A M E W V C D H D C V	369	3541	ACAGCTTATGGAAATGCTCTTTATCTCTATCAGAAAGAGACATAGATTTAGGCT	3600
1201	GATAAGTCGAGGAGTCACTGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	1260	3601	GATTAATATCTCTACAGCTTTTGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	3660
370	D X S D E V N C S C H S Q G L V E C R N	389	3661	GTTAATCTGGAGCTGCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTT	3720
1261	GGACATGATTCACCCAGCTTTCAATGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	1320	3721	ACACTTGGAAATTTAGGCTACAGCAGCAACGTCGAGGTAGTATCATATGATATCATATG	3780
390	G Q C I F S T G D G D E D C K D G S	409	3781	TGCCATGTGGTGTGTGTAACCCAGTAACTGCTCATTTGATTATTAAGAGCCAAGATAA	3840
1321	GATGAGGAGAACTGACAGCTGATTCAGATTCATGTCAGAGGAGGACCAAAAGATGGCTC	1380	3841	TTTACATGTTTAAAGTATTACTATTACCCCTCTCTAAATGTTGCAATATCTGAGAACT	3900
410	D E E N C S V I Q T S C Q E G D R O R C L	429	3901	GATAAAGACAGCAATAAAGACCACTGCTCATCAATTTAGGTAGCAGACATATTGAATG	3960
1381	TACAATCCCTGCTTGTATGATGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	1440	3961	CAAGTCTTTAGATATCAATATTAACTTGAACCTTGACATTTAGGACCCCACTCTGCTGATG	4020
430	Y N P C L D S C G G S S L C D P H N S L	449	4021	ATATCAAGATCATATTTATAGAGAGCTCTATAGAACTGCTCTCATAGCTGGGTTTG	4080
1441	AATACTGTAGTCAATGTGAACCAATTTAATTTGGAATCTGCTGATGAATTTGGCTACACA	1500	4081	TTACAGATATGAGTTGGCTGATTTGAGAGTGCACAACTACATCTATATTTATGGGCA	4140
450	N N C S Q C E P I T L E L C N N L P Y N	469	4141	TATTTCTTTTACTATGTGGCAAGCTGCATTAATCTTGCAAGAGAGCAATTTAG	4200
1501	AGTCAAGATTCACAAATTTTGTGGCAGAGACTCAAAAGGAGCATTCATCAGCTGG	1560	4201	ATGAGAGATGCAATTTTAAAGAAATTAATTTGCAATCCCTCTCTTAAATTAATTTA	4260
470	S T S Y P N Y C F G H R T Q K E A S I G W	489	4261	TTTTTCAAGTTTCTGGCTTCATCCATCCAAAGTCAATAAGAGCATATTTTAGAGC	4320
1561	GAGTCTCTCTTTTCCCTGCACTTGTTCACCACTGTATATAATCTCATCTCTCTCTCT	1620	4321	ACAGTAAAGCTTTGCTGAGTAAACATTTTGAATTTTCTCTCAAAAGATGTTTAAAT	4380
490	E S S L F P A L V Q T N C Y K Y L M F F	509	4381	CTGGTTCTCTCTCATTTGGTAATTAATAATTTAGAAATGATTTTAGCTCTAGGCCACTTT	4440
1621	CTTGTGACCAATTTTGTGTAAGATGTGTAATACAGGCGAGGATATCCCTCTCTG	1680	4441	ACCGAATCAATTTCTGAAGCAATTTAGTGGTAAAGATGATTTTCCCACTAAAGAACTT	4500
510	S C T L L V P K C D V N T G E R I P P C	529	4501	TAAACACAAATCTTCATATATCTTAAATTTAGTCAGGATCATTTTGGCTTTTA	4560
1681	AGGGCATCTGTGAACACTCTAAAGAACGCTGTGAGTCTGTTCTTGGGATTTGGGCTTA	1740	4561	ACCACTAGGATTTCCCTACTAATCTCCAGCAGCACTGGAGCTGCTGCAATTTCAAT	4620
530	R A L C E S H E R C E S L L G I V G L	549	4621	AGATCTACTGCTCAATTTTATACATGATTTTGTATCTTTCTCTGTTGTAACATCAT	4680
1741	CAGTGGCTGAGACAGATTCAGTCAATTTCCAGAGGAAATTCAGACAAATCAAACT	1800	4681	GAAATCAAAAGTGTAGCAATTTCTATCTATCTCTCTCTCTCTCTCTCTCTCTCTCTCT	4740
550	Q W P E D T D C S Q F P E E N S D N O T	569	4741	ACCTTAAAGAGAGTGTGAAATCCAGCACTGAATTTGGTGTGAGTGTGTAAGATTTCA	4800
1801	TGCTGATGCTGATGAATATGTGGAAGATGCTCACTGATGCTATTTCAAGTCCGCTCA	1860	4801	AGAACTATGTCAGTTTGTGACGTGTGAGTACATCTCAATGTATCACTTTTATGCT	4860
570	C L M P D E Y V E E C S P S H F K C R S	589	4861	TGCTCACTTGGCTCAGTGAATATATATATATCTATTTTAAATTAATTTCTTAAATC	4920
1861	GGACAGTGTGTTCTGGCTTCCAGAAATGTGATGCCAGGCGCTGAGCATGACAGT	1920	4921	<u>AAATTAAT</u> TGTA	4933
590	G Q C V L A S R D G Q A D C D D D S	609			
1921	GATGAGGAAATCTGTTGTGTAAGAGAGATCTTTGGAACTGTCATCAATCAAAACA	1980			
610	D E E N C G K E R D L W E C P S N K O	629			
1981	TGTTGAAGCACAGTGTCTGCGATGGGTCCGAGCTGCTGATTCATGGACGAG	2040			
630	C L K H T V I C D G F P D C F D Y H D E	649			
2041	AAAACTGCTCATTTTGGCAAGTCAATGAGCTGGAATGTGCAAAACCTGGTGTGTCTCA	2100			
650	K H C S F C Q D D E L E C A N H A C V S	669			
2101	CTGACCTGTGGTGTGATGGTGAAGCGAGTCTCAGACAGTTCAGATGAATGGGAGCT	2160			
670	R D L W C D G E A D C S D S D E W D C	689			
2161	GTGACCTCTCTATAAATGTGAACCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	2220			
690	V T L S I N V H S S S F L M Y H R A A T	709			
2221	GAACACATGTGTGTCAGATGGCTGGCGAGAGATTTGAGTCACTGGCTGCAAGCAG	2280			
710	B H H V C A D G W Q E I L S O L A C K Q	729			

FIG. 1. Nucleotide sequence of human corin cDNA and its deduced amino acid sequence. The potential codon for the initial methionine, the translation stop codon, and the polyadenylation signal were in **bold-face type** and underlined. The putative transmembrane domain was double underlined. The 19 potential N-linked glycosylation sites are in **boldface type** and double underlined. An arrow indicates the putative cleavage site for the activation of the serine protease. The active site residues of the catalytic triad (His⁸⁴³, Asp⁸⁹², and Ser⁸⁸⁵) are in **boldface type** and underlined.

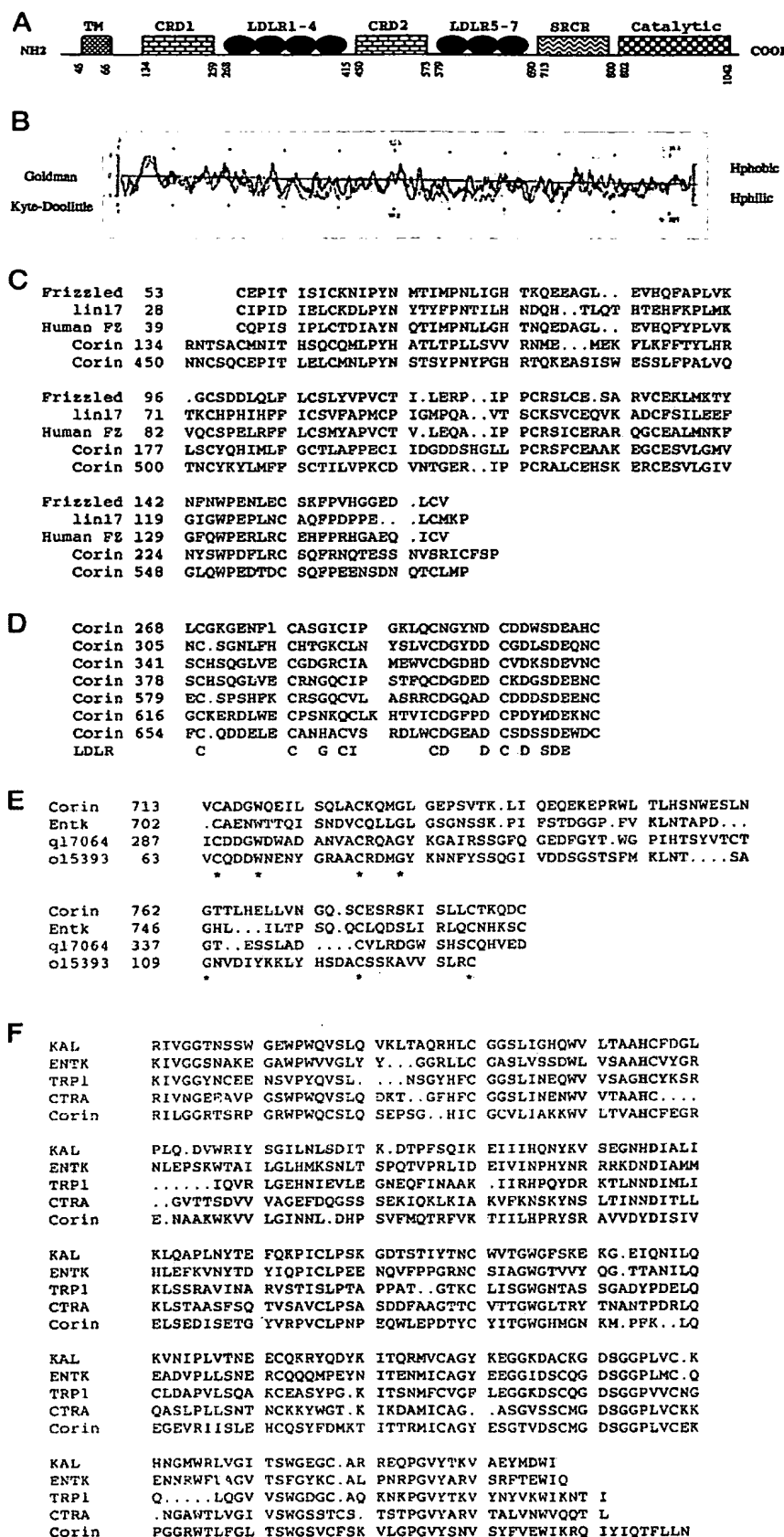


FIG. 2. A, a schematic presentation of the domain structure of corin protein. The transmembrane domain (TM), frizzled-like cysteine-rich domains (CRD), LDL receptor repeats (LDLR), scavenger receptor cysteine-rich domain (SRCR), and serine protease catalytic domain (Catalytic) are indicated. Numbers correspond to the amino acid residues of the ORF shown in Fig. 1. B, hydropathy plots of the deduced amino acid sequence of corin by Goldman and Kyte-Doolittle methods, respectively (36). *Hphobic*, hydrophobic; *Hphilic*, hydrophilic. C, alignment of amino acid sequences of the frizzled-like cysteine-rich domains from corin and other members of the frizzled family, including Frizzled in *Drosophila*, lin-17 in *C. elegans*, and FZ-1 in human. D, alignment of amino acid sequences of the seven LDL receptor repeats of corin with the consensus sequence derived from the human LDL receptor. E, alignment of amino acid sequences of the scavenger receptor-like cysteine-rich domains from corin and human enterokinase (*Entk*), sea urchin speract receptor (*q17064*) and human scavenger receptor I (*o15393*). Asterisks indicate conserved residues. F, alignment of amino acid sequences of protease domains from human corin, prekallikrein (KAL), enterokinase (ENTK), trypsin (TRP1), and bovine chymotrypsinogen A (CTRA).

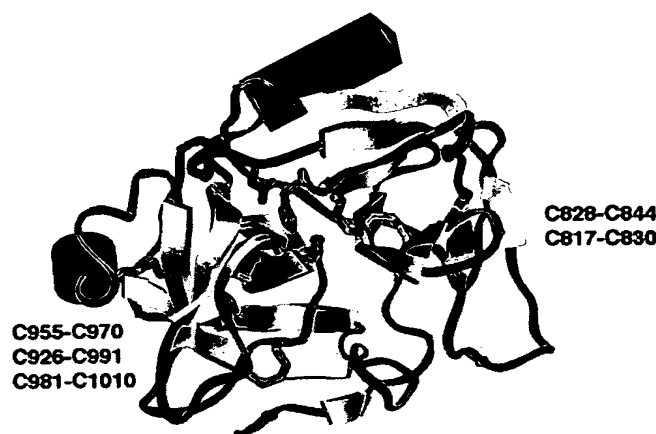


FIG. 3. Molecular model of the protease domain of corin between amino acids 802 and 1042. A corin model was built based on the structure of bovine chymotrypsinogen A, as described under "Experimental Procedures." The active site residues of the catalytic triad (His⁸⁴³, Asp⁸⁹², and Ser⁹⁸⁵) are shown in purple. Four disulfide bonds in the corin model (Cys⁸²⁸-Cys⁸⁴⁴, Cys⁹⁵⁵-Cys⁹⁷⁰, Cys⁹²⁶-Cys⁹⁹¹, and Cys⁹⁸¹-Cys¹⁰¹⁰) that correspond to the disulfide bonds in the catalytic domain of chymotrypsinogen (Cys⁴²-Cys⁶⁸, Cys¹⁶⁸-Cys¹⁸², Cys¹³⁶-Cys²⁰¹, and Cys¹⁹¹-Cys²²⁰) are shown in blue. The side chains of Cys⁸¹⁷ and Cys⁸³⁰ of the corin model are in an acceptable proximity to form a disulfide bond (pink). The distance between the C- α atoms from the chymotrypsinogen template (Val³¹ and Gly⁴⁴) corresponding to these two cysteine residues is 5.08 Å, and the distance between the sulfur atoms after rotamer searching of the cysteine side chains is about 2.5 Å. The potential disulfide bond between Cys⁷⁹⁰ and Cys⁹¹² of corin corresponding to the disulfide bond between Cys¹ and Cys¹²² of chymotrypsinogen is not included in the model.

This region contains 88 amino acids and is homologous to the cysteine-rich motif found in the macrophage scavenger receptor (30). This motif is also present in the sea urchin spermatozoa speract receptor (31, 32) and the vertebrate serine protease, enterokinase (27).

At the carboxyl terminus of corin protein between amino acid residues 802 and 1042, there is a trypsin-like serine protease domain (Fig. 2A). This protease domain is highly homologous to the catalytic domain of members of the trypsin superfamily. For example, amino acid sequence identities between corin and prekallikrein (33), factor XI (34), and hepsin (35) are 40, 40, and 38%, respectively. All essential features of serine protease sequences are well conserved in corin (Figs. 1 and 2F). The active site residues of the catalytic triad are located at His⁸⁴³, Asp⁸⁹², and Ser⁹⁸⁵. The amino acid residues forming the substrate specificity pocket are located at Asp⁹⁷⁹, Gly¹⁰⁰⁷, and Gly¹⁰¹⁸. These residues are predicted to bind the substrate P1 residues, suggesting that corin would cleave its substrate after basic residues, such as lysine or arginine. In addition, a putative activation cleavage site was found at Arg⁸⁰¹, suggesting that corin would be synthesized as an inactive zymogen and that another trypsin-like enzyme was required for its activation.

In the protease domain, there are 12 cysteine residues. Potential pairing of these cysteine residues can be predicted by comparing with other well studied serine proteases, such as trypsin and chymotrypsin. First three pairs of cysteine residues present in essentially all members of the trypsin superfamily are located at Cys⁸²⁸-Cys⁸⁴⁴, Cys⁹⁵⁵-Cys⁹⁷⁰, and Cys⁹⁸¹-Cys¹⁰¹⁰. Two more pairs of cysteine residues are present at the positions Cys⁷⁹⁰-Cys⁹¹² and Cys⁹²⁶-Cys⁹⁹¹. These two pairs of cysteine residues are commonly found in a subfamily of two-chain serine proteases, such as chymotrypsin and prekallikrein (33). The presence of Cys⁷⁹⁰ and Cys⁹¹² indicated

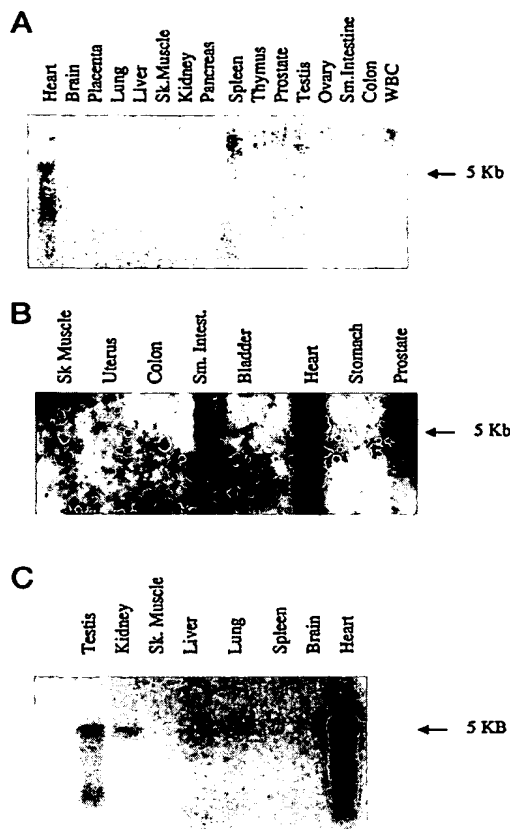


FIG. 4. Northern analysis of corin mRNA expression. Human and mouse multiple tissue Northern blots were hybridized with human and mouse corin cDNA probes, respectively. In human tissues (A and B), corin mRNA was detected only in samples from heart. In mouse tissues (C), abundant expression of corin mRNA was detected in samples from heart. Weak signals were also detected in samples from testis and kidney.

that, after the activation cleavage at Arg⁸⁰¹, the catalytic domain of corin would remain attached to the rest of molecule by a disulfide bond. Interestingly, there is one additional pair of cysteine residues, Cys⁸¹⁷ and Cys⁸³⁰, present in corin. Cysteine residues at these two positions were not found in any other serine proteases in vertebrates. A search of data bases showed that a chymotrypsinogen-like serine protease from the lugworm, *Arenicola marina*, had two cysteine residues at the corresponding positions.² A model of the corin protease domain was built based on the structure of bovine chymotrypsinogen A (Fig. 3). Based on this corin model, where the C- α atoms of these two cysteine residues were held fixed during energy minimization, the distance between the sulfur atoms of their side chains is about 2.5 Å after rotamer searching. The model indicates that these two cysteines are likely to form a disulfide bond connecting two β -sheets in the core of the protease domain (Fig. 3).

Northern Analysis of Corin mRNA Expression—To determine expression of the corin gene in human tissues, Northern hybridization was performed using human corin cDNA probes. As shown in Fig. 4A, an ~5-kb transcript was detected only in the heart but not in other tissues including brain, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, colon, and leukocytes. Since the heart is mainly composed of cardiac muscles, Northern analysis was

² J. Eberhardt, GenBank™ accession number G1160388.

FIG. 5. Analysis of corin mRNA expression by *in situ* hybridization in an adult mouse heart. Tissue sections from atrium (B) and ventricle (A) were stained with hematoxylin/eosin. Corin mRNA was detected by *in situ* hybridization using a mouse corin cDNA probe. Expression of corin mRNA was found in the cardiac myocytes of both the atrium (D) and the ventricle (C) as shown by white spots.

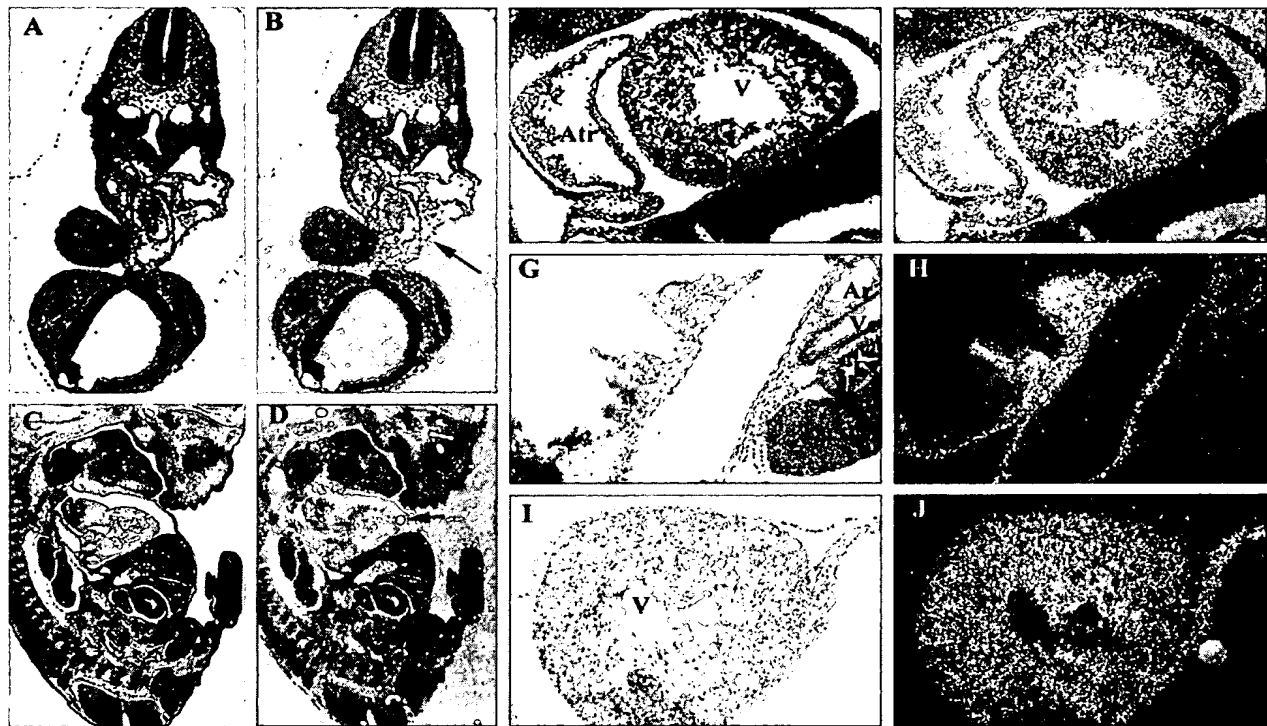
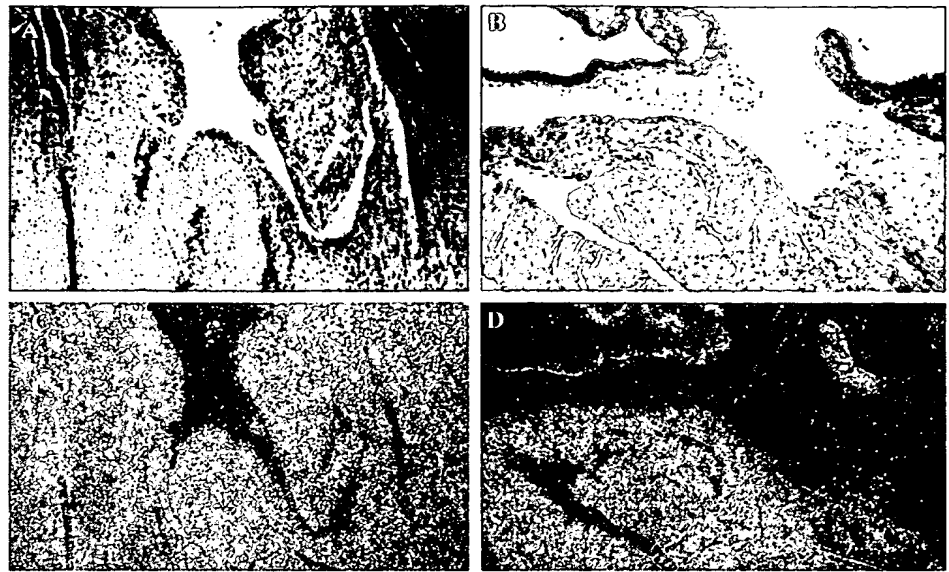


FIG. 6. Expression of corin mRNA in the developing heart. Tissue sections were prepared from mouse embryos at day E9.5 (A and B), E11.5 (C and D), E12.5 (E and F), and E15.5 (G–J) and stained with hematoxylin/eosin (A, C, E, G, and I). Corin mRNA expression was detected by *in situ* hybridization in developing heart by E9.5 (B) and E11.5 (D) as indicated by arrows. The expression was prominent in the primary atrial septum and the trabecular ventricular compartment by E12.5 (F). By E15.5, corin mRNA was detected in most cardiac myocytes in both atrium (H) and ventricle (J). Abbreviations used in E, G, and I are as follows: Atr, atrium; V, ventricle; Ar, aorta; Vc, vena cava; E, esophagus; Lu, lung.

performed to examine the presence of corin mRNA in other human muscle-rich tissues. Again, corin mRNA was detected in the heart but not in uterus, small intestine, bladder, stomach, and prostate (Fig. 4B).

To examine corin mRNA expression in mice, the full-length mouse corin cDNA was cloned by a PCR-based strategy. Mouse corin cDNA shared 89% sequence identities with human corin cDNA (data not shown). Northern analysis was performed with RNA samples from mouse tissues. As shown in Fig. 4C, a prominent transcript of ~5 kb was detected in samples derived

from the heart. In contrast to Northern analysis with human samples, low levels of corin mRNA were also detected in samples derived from the testes and kidneys.

Mouse Corin mRNA Expression in Adult and Embryonic Hearts—*In situ* hybridization was performed to determine the temporal and special expression of corin mRNA. In adult mice (Fig. 5), corin mRNA was detected in cardiac myocytes of both atrium and ventricle. The level of expression appeared to be higher in the atrium than the ventricle. During embryonic development, corin mRNA was first detected at E9.5 in both

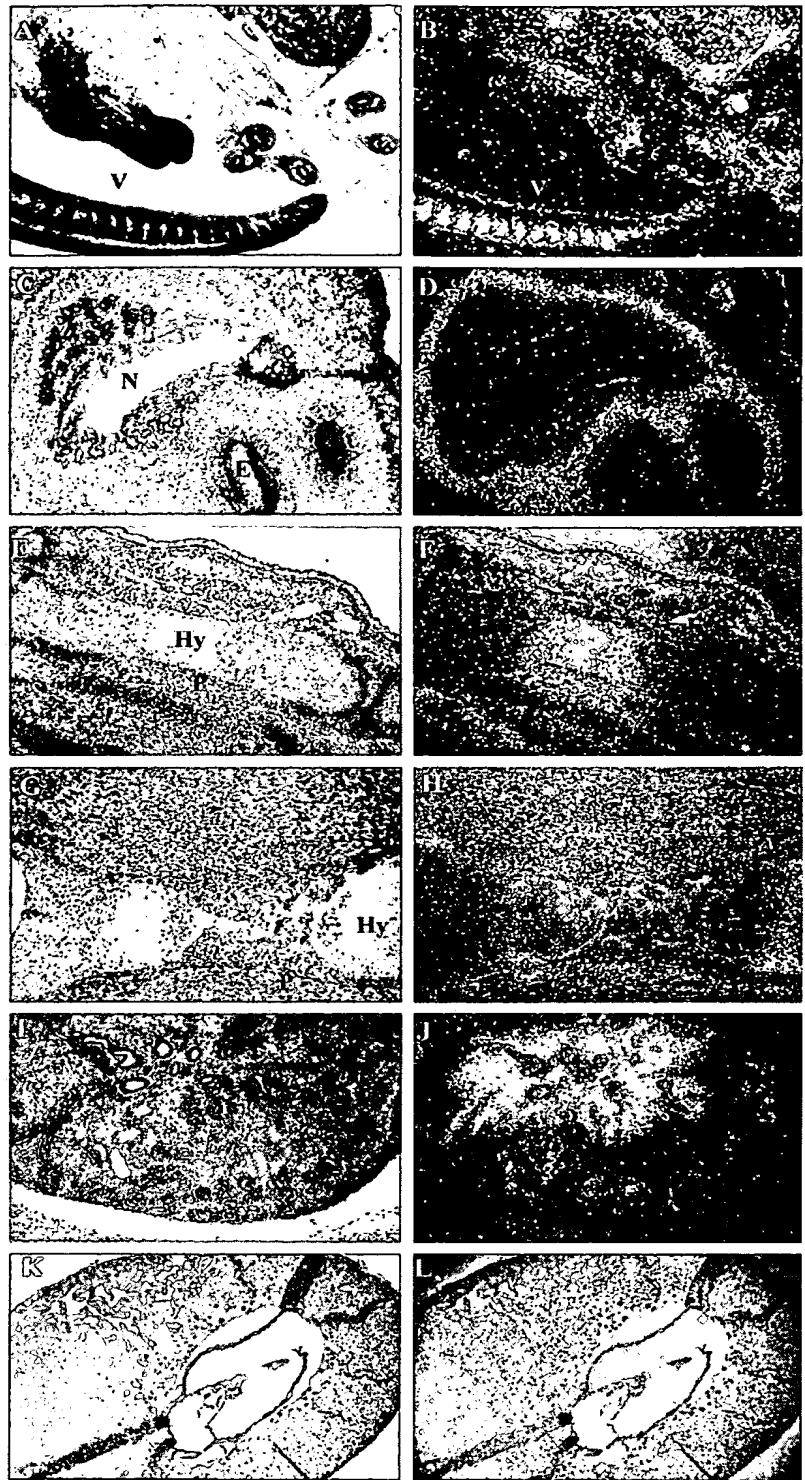


FIG. 7. Expression of corin mRNA in other tissues during embryonic development. Tissue sections were stained with hematoxylin/eosin. *In situ* hybridization was performed using a mouse corin cDNA probe, as described under "Experimental Procedures." *A* and *B*, expression of corin mRNA in cartilage primordia of vertebral bodies of an E13.5 embryo. *C* and *D*, expression of corin mRNA in the turbinates primordia around the nasal and eye cavities of an E15.5 embryo. *E* and *F*, expression of corin mRNA in a developing digital bone in a front paw at E15.5. Corin mRNA was detected in the region adjacent to the hypertrophic chondrocytes and in the perichondrocytes. *G* and *H*, in a more matured digital bone in a hind limb of an E15.5 embryo, a similar pattern of corin mRNA expression was found in the region adjacent to the hypertrophic chondrocytes and in the perichondrocytes. *I* and *J*, expression of corin mRNA in the medulla of a developing kidney at E15.5. *K* and *L*, expression of corin mRNA in the decidua of a pregnant uterus. Abbreviations used are: V, vertebral bodies; N, nasal cavity; E, eye cavities; Hy, hypertrophic chondrocytes; P, perichondrocytes.

atrium and ventricle of the developing heart (Fig. 6B). Between E11.5 and E13.5, corin mRNA was highly expressed in the thickened atrial wall and in the regions that underwent trabeculation in the ventricle (Fig. 6, D and F). By E15.5, corin mRNA in the heart was more abundant, especially in primary atrial septa (Fig. 6H). Weak signals appeared to be present in developing aorta and vena cava but not in the esophagus and lungs (Fig. 6H). The expression of corin mRNA in the heart was

maintained in the subsequent embryonic stages (not shown).

Corin mRNA Expression in Other Tissues—In addition to the heart, corin mRNA was also detected in other mouse tissues by *in situ* hybridization. For example, corin mRNA was present in the uterus of pregnant mice and in the developing kidneys. In the uterus (Fig. 7L), corin mRNA expression was most abundant in the decidua cells close to the implantation site of the embryo. In the developing kidneys at E15.5, corin mRNA was

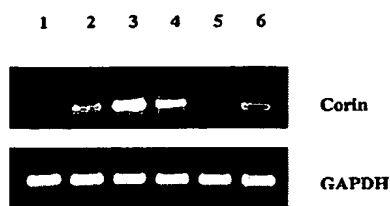


FIG. 8. Analysis of corin mRNA expression in tumor cell lines by RT-PCR. RNA samples were isolated from human tumor cell lines. RT-PCR experiments were performed using oligonucleotide primers derived from human corin cDNA. Corin mRNA was detected in samples from Hec-1-A, U2-OS, SK-LMS-1, RL95-2, and AN3-CA cells (upper panel, lanes 2–6) but not in samples from HeLa cells (upper panel, lane 1). In a control experiment, PCR reactions were performed with specific oligonucleotide primers for the human GAPDH gene. GAPDH mRNA was detected in samples from all cell lines (lower panel, lanes 1–6).

highly expressed in the stromal cells in the medulla but not in the cortex of the kidney (Fig. 7J). This finding was consistent with the results of Northern analysis in which a corin transcript was found in RNA samples from mouse kidneys (Fig. 3C).

Interestingly, *in situ* hybridization also identified corin mRNA in several cartilage-derived structures, such as the vertebra in the tail, the turbinate in the head, and the long bones in the limbs (Fig. 7, B, D, F, and H). Fig. 7B showed the expression of corin mRNA in cartilage primordia of vertebral bodies in the posterior of an E13.5 embryo. By E15.5, the level of corin mRNA expression in the vertebra was much lower as the vertebra became more matured (data not shown), indicating that corin may play a role in the differentiation of chondrocytes. This notion was supported by the expression of corin mRNA in developing limbs. Fig. 7, E and F, showed an early developing digital bone that consisted of three types of cells as follows: hypertrophic chondrocytes at the center, prehypertrophic chondrocytes next to the hypertrophic zone, and proliferating chondrocytes at the both ends. Corin mRNA was found mostly in the prehypertrophic chondrocytes (Fig. 7F). Hybridization signals were also present in perichondrium (Fig. 7F). Fig. 7, G and H, showed a long bone in a hind limb that was at a more advanced developmental stage. The central hypertrophic zone was replaced by vascularized tissues containing bone marrow cells and osteoblasts. Nevertheless, similar expression pattern of corin mRNA was found in the narrow zone of the prehypertrophic chondrocytes and in the perichondrium. These results indicated that corin expression was associated with a specific stage of chondrocyte differentiation.

Corin mRNA Expression in Human Tumor Cell Lines—A number of human cancer cell lines were screened by Northern and RT-PCR analyses for the presence of corin mRNA. In most cell lines, such as HL60, HeLa, K562, MOLT-4, RAJI, SW480, A549, and G36, corin mRNA was undetectable (data not shown). However, corin mRNA was found in several cell lines derived from uterus tumors or osteosarcoma. As shown in Fig. 8, corin mRNA was detected by RT-PCR in endometrium carcinoma cell lines HEC-1-A, AN3 CA, and RL95-2, leiomyosarcoma cell line SK-LMS-1, as well as in osteosarcoma cell line U2-OS. The result is consistent with the finding by *in situ* hybridization in which corin mRNA was highly expressed in the developing bones in embryos as well as in the maternal uterus.

Chromosomal Localization of the Human Corin Gene—FISH analysis was performed to determine the chromosomal locus of the human corin gene. Specific fluorescent spots were found at 4p12-13, a region adjacent to the centromere on the short arm of chromosome 4 (Fig. 9). The result was confirmed in a subsequent experiment in which a genomic probe previously mapped to 4p15.3 was co-localized with the corin gene probe (data not

shown). A search of the OMNI human genetic data base indicated that a congenital heart disease locus, total anomalous pulmonary venous return (TAPVR), was previously mapped to this region at 4p13-q12 (37).

DISCUSSION

In this study, we describe the cloning and initial characterization of a novel cDNA from the human heart that encodes a putative transmembrane serine protease, which we have designated as corin. The presence of a hydrophobic transmembrane domain at its amino terminus and the absence of a signal peptide suggest that corin is a type II transmembrane protein. In the extracellular region of corin, there is a trypsin-like catalytic domain that contains all conserved structural features of serine proteases, such as the catalytic triad, the activation cleavage site, the substrate specificity pocket, and the essential cysteine residues. Interestingly, the protease domain of corin contains two unique cysteine residues, Cys⁸¹⁷ and Cys⁸³⁰, that are not present in other trypsin-like serine proteases in vertebrates. Molecular modeling showed that these two cysteine residues are likely to form a disulfide bond connecting two β -sheets in the core of the protease domain (Fig. 3). A search of genomic data bases showed that a chymotrypsin-like protease found in the lugworm, *A. marina*, also has two cysteine residues at the corresponding positions. It is not clear whether these two cysteine residues are maintained through a convergent or divergent evolution. Nevertheless, the presence of such an unusual pair of cysteine residues in both corin and the lugworm protease suggests an important biological function of the disulfide bond. One potential possibility is that the disulfide bond may contribute to stability of the proteases.

Although members of the trypsin superfamily are known to contain a variety of domain structures such as kringle and epidermal growth factor-like domains that are important for protein-protein interactions, this is the first report of the presence of a frizzled-like cysteine-rich domain in this extended family. Originally, the frizzled gene was identified in *Drosophila* (38). The gene encodes a seven-transmembrane receptor that is required for proper development of hairs, bristles, and ommatidia of the fruit fly (19, 39). Later, other Frizzled proteins have been identified in many other species. They all contain a well conserved extracellular cysteine-rich domain and a seven-transmembrane domain and act as receptors for secreted Wnt glycoproteins (for review see Refs. 40 and 41). The cysteine-rich domain, which is about 120 amino acids in length and contains a motif of 10 invariantly spaced cysteine residues, has been shown to be necessary and sufficient for the binding of the Wnt ligands (20, 42). Recent studies demonstrated that Frzb, a secreted frizzled-like protein without the seven-transmembrane domain, is expressed in the Spemann organizer of frog embryos and can bind and inhibit Wnt-8 (43, 44). In addition, similar frizzled-like cysteine-rich domains have also been found in several other proteins, including mouse collagen (XVIII) α 1 chain (45), human carboxypeptidase Z (46), and several receptor tyrosine kinases (47–49). The function of the cysteine-rich domain in these proteins has not been determined. Corin is unique in that it contains the frizzled-like cysteine-rich domains and a serine protease domain. The presence of frizzled-like domains in corin implies that corin may play an important role in development by directly interacting with Wnt proteins.

The temporal and special pattern of corin gene expression further supported a potential developmental function of corin. In mice, corin mRNA was detected in the cardiac myocytes of the embryonic heart as early as E9.5 (Fig. 6B). The expression was most prominent in the primary atrial septum and the trabecular ventricular compartment by E11.5–13.5 (Fig. 6, D



FIG. 9. Chromosomal localization of the human corin gene by FISH. A fluorescent-labeled genomic DNA probe containing the human corin gene was hybridized to metaphase chromosomes derived from PHA-stimulated peripheral blood lymphocytes. Hybridization signals are shown as bright blue spots and indicated by white arrows (left panel). The position of the corin locus on human chromosome 4 is illustrated in a diagram (right panel).

and *F*). During this period, an active process of looping and remodeling takes place in the embryonic heart. As a result, outflow tracts are formed, and the original single tube-like heart is reorganized into a four-chambered structure. Growth factors, such as bone morphogenic proteins and the transforming growth factor- β family members, are known to play a critical role during the embryonic heart development (50). Recent studies in *Drosophila* showed that the *wingless* (*wg*) gene, a homologue of the *wnt* oncogene in mammals, is directly involved in heart formation (51). It has been suggested that similar signaling pathways also contributed to the heart development in vertebrate (52). It is possible that corin could participate in such developmental pathways by interacting directly with Wnt proteins or other growth factors.

In addition to the heart, corin mRNA was identified in other tissues, such as the pregnant uterus and developing kidneys and bones. The expression of corin mRNA in these tissues appeared to be cell type-specific. For example, in developing long bones corin mRNA was specifically expressed in the prehypertrophic chondrocytes. It is known that skeletal bones are derived from two different processes, intramembranous and endochondral ossification. In the former case, mesenchymal tissues are directly converted into bones, whereas in the latter case the mesenchymal cell is converted to bone via cartilage as an intermediate step. The vertebrae, long bones, and certain fragments of skull are formed by endochondral ossification (53). In these bones, mesenchymal cells first become chondrocytes that in turn differentiate from proliferating chondrocytes to prehypertrophic chondrocytes and finally to hypertrophic chondrocytes. The hypertrophic chondrocytes eventually undergo apoptosis followed by vascularization and ossification. This process of chondrocyte differentiation has been shown to be tightly regulated by hedgehog proteins, bone morphogenic proteins, and parathyroid hormone-related protein (54–57). The specific expression of corin mRNA in a subset of chondrocytes indicated that corin may also be involved in this cell differentiation process.

Finally, by FISH analysis the human corin gene was located on the short arm of chromosome 4 (4p12-13) (Fig. 9). A search of the OMNI human genetic data base showed that a disease

locus, total anomalous pulmonary venous return (TAPVR), had been previously mapped to this region. TAPVR is a rare cyanotic form of congenital heart defects in which the pulmonary vein connected abnormally to the right atrium or one of the venous tributaries instead of the left atrium. The molecular mechanism responsible for this developmental defect in the heart is unknown. A linkage study of a large Utah-Idaho family that included 14 affected individuals localized the TAPVR locus to a 30-centimorgan interval on 4p13-q12 (37). The findings that the corin gene and the TAPVR locus are co-localized on chromosome 4 and that corin mRNA is highly expressed in the embryonic heart, particularly in the region where outflow tracts were formed, suggest that corin is an attractive candidate for the TAPVR gene. The isolation of the corin cDNA provided a useful tool to study further this intriguing possibility.

Acknowledgments—We thank Drs. W. Dole and G. Rubanyi for their encouragement and helpful discussions.

REFERENCES

1. Neurath, H. (1984) *Science* 224, 350–357
2. Huber, R., and Bode, W. (1978) *Acc. Chem. Res.* 11, 114–122
3. Davie, E. W., Fujikawa, K., and Kisiel, W. (1991) *Biochemistry* 30, 10363–10370
4. Morisato, D., and Anderson, K. V. (1995) *Annu. Rev. Genet.* 29, 371–399
5. LeMosy, E. K., Kemler, D., and Hashimoto, C. (1998) *Development* 125, 4045–4053
6. Konrad, K. D., Goralski, T. J., Mahowald, A. P., and Marsh, J. L. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 6819–6824
7. Anderson, K. V., Schneider, D. S., Morisato, D., Jin, Y., and Ferguson, E. L. (1992) *Cold Spring Harbor Symp. Quant. Biol.* 57, 409–417
8. Tsuji, A., Torres-Rosado, A., Arai, T., Le Beau, M. M., Lemons, R. S., Chou, S. H., and Kurachi, K. (1991) *J. Biol. Chem.* 266, 16948–16953
9. Torres-Rosado, A., O'Shea, K. S., Tsuji, A., Chou, S. H., and Kurachi, K. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 7181–7185
10. Wu, Q., Yu, D., Post, J., Halks-Miller, M., Sadler, J. E., and Morser, J. (1998) *J. Clin. Invest.* 101, 321–326
11. Appel, L. F., Prout, M., Abu-Shumays, R., Hammonds, A., Garbe, J. C., Fristrom, D., and Fristrom, J. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 4937–4941
12. Yamaoka, K., Masuda, K., Ogawa, H., Takagi, K., Umamoto, N., and Yasuoka, S. (1998) *J. Biol. Chem.* 273, 11895–11901
13. Paoloni-Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C., and Antonarakis, S. E. (1997) *Genomics* 44, 309–320
14. Jen, Y., Manova, K., and Benzeval, R. (1997) *Dev. Dyn.* 208, 92–106
15. Wang, D., Bode, W., and Huber, R. (1985) *J. Mol. Biol.* 185, 595–624
16. Ponder, J. W., and Richards, F. M. (1987) *J. Mol. Biol.* 193, 775–791
17. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr., Brice, M. D.,

- Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535-542
18. Hartmann, E., Rapoport, T. A., and Lodish, H. F. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 5786-5790
19. Vinson, C. R., Conover, S., and Adler, P. N. (1989) *Nature* **338**, 263-264
20. Bhanot, P., Brink, M., Samos, C. H., Hsieh, J. C., Wang, Y., Macke, J. P., Andrew, D., Nathans, J., and Nusse, R. (1996) *Nature* **382**, 225-230
21. Sawa, H., Lobel, L., and Horvitz, H. R. (1996) *Genes Dev.* **10**, 2189-2197
22. Chan, S. D., Karpf, D. B., Fowlkes, M. E., Hooks, M., Bradley, M. S., Vuong, V., Bambino, T., Liu, M. Y., Arnaud, C. D., Strewler, G. J., and Nissenson, R. A. (1992) *J. Biol. Chem.* **267**, 25202-25207
23. Brown, M. S., Herz, J., and Goldstein, J. L. (1997) *Nature* **388**, 629-630
24. Krieger, M., and Herz, J. (1994) *Annu. Rev. Biochem.* **63**, 601-637
25. Kounnas, M. Z., Chappell, D. A., Strickland, D. K., and Argraves, W. S. (1993) *J. Biol. Chem.* **268**, 14176-14181
26. Catterall, C. F., Lyons, A., Sim, R. B., Day, A. J., and Harris, T. J. (1987) *Biochem. J.* **242**, 849-856
27. Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W., and Sadler, J. E. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91**, 7588-7592
28. Schonbaum, C. P., Lee, S., and Mahowald, A. P. (1995) *Proc. Natl. Acad. Sci. U. S. A.* **92**, 1485-1489
29. Hong, C. C., and Hashimoto, C. (1995) *Cell* **82**, 785-794
30. Matsumoto, A., Naito, M., Itakura, H., Ikemoto, S., Asaoka, H., Hayakawa, I., Kanamori, H., Aburatani, H., Takaku, F., Suzuki, H., Kobari, Y., Miyai, T., Takahashi, K., Cohen, E. H., Wydro, R., Housman, D. E., and Kodama, T. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87**, 9133-9137
31. Thorpe, D. S., and Garbers, D. L. (1989) *J. Biol. Chem.* **264**, 6545-6549
32. Dangott, L. J., Jordan, J. E., Bellet, R. A., and Garbers, D. L. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 2128-2132
33. Chung, D. W., Fujikawa, K., McMullen, B. A., and Davie, E. W. (1986) *Biochemistry* **25**, 2410-2417
34. Fujikawa, K., Chung, D. W., Hendrickson, L. E., and Davie, E. W. (1986) *Biochemistry* **25**, 2417-2424
35. Leytus, S. P., Loeb, K. R., Hagen, F. S., Kurachi, K., and Davie, E. W. (1988) *Biochemistry* **27**, 1067-1074
36. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105-132
37. Bleyl, S., Nelson, L., Odelberg, S. J., Ruttenberg, H. D., Otterud, B., Leppert, M., and Ward, K. (1995) *Am. J. Hum. Genet.* **56**, 408-415
38. Gubb, D., and Garcia-Bellido, A. (1982) *J. Embryol. Exp. Morphol.* **68**, 37-57
39. Zheng, L., Zhang, J., and Carthew, R. W. (1995) *Development* **121**, 3045-3055
40. Cadigan, K. M., and Nusse, R. (1997) *Genes Dev.* **11**, 3286-3305
41. Shulman, J. M., Perrimon, N., and Axelrod, J. D. (1998) *Trends Genet.* **14**, 452-458
42. Lin, K., Wang, S., Julius, M. A., Kitajewski, J., Moos, M., Jr., and Luyten, F. P. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94**, 11196-11200
43. Wang, S., Krinks, M., Lin, K., Luyten, F. P., and Moos, M., Jr. (1997) *Cell* **88**, 757-766
44. Leys, L., Bouwmeester, T., Kim, S. H., Piccolo, S., and De Robertis, E. M. (1997) *Cell* **88**, 747-756
45. Rehn, M., and Pihlajaniemi, T. (1995) *J. Biol. Chem.* **270**, 4705-4711
46. Song, L., and Fricker, L. D. (1997) *J. Biol. Chem.* **272**, 10543-10550
47. Xu, Y. K., and Nusse, R. (1998) *Curr. Biol.* **8**, R405-R406
48. Masiakowski, P., and Yancopoulos, G. D. (1998) *Curr. Biol.* **8**, R407
49. Saldanha, J., Singh, J., and Mahadevan, D. (1998) *Protein Sci.* **7**, 1632-1635
50. Wu, X., Golden, K., and Bodmer, R. (1995) *Dev. Biol.* **169**, 619-628
51. Park, M., Wu, X., Golden, K., Axelrod, J. D., and Bodmer, R. (1996) *Dev. Biol.* **177**, 104-116
52. Bodmer, R., and Venkatesh, T. V. (1998) *Dev. Genet.* **22**, 181-186
53. Gilbert, S. F. (1994) *Developmental Biology* (Gilbert, S. F., ed) 4th Ed., Sinauer Associates, Inc., Sunderland, MA
54. Lanske, B., Karaplis, A. C., Lee, K., Luz, A., Vortkamp, A., Pirro, A., Karperien, M., Defize, L. H. K., Ho, C., Mulligan, R. C., Abou-Samra, A. B., Juppner, H., Segre, G. V., and Kronenberg, H. M. (1996) *Science* **273**, 663-666
55. Storm, E. E., Huynh, T. V., Copeland, N. G., Jenkins, N. A., Kingsley, D. M., and Lee, S. J. (1994) *Nature* **368**, 639-643
56. Vortkamp, A., Lee, K., Lanske, B., Segre, G. V., Kronenberg, H. M., and Tabin, C. J. (1996) *Science* **273**, 613-622
57. Zou, H., Wieser, R., Massague, J., and Niswander, L. (1997) *Genes Dev.* **11**, 2191-2203

Exhibit 45

BIOCHEMISTRY

Coordinating Author GEOFFREY ZUBAY
COLUMBIA UNIVERSITY



ADDISON-WESLEY PUBLISHING COMPANY

READING, MASSACHUSETTS ♦ MENLO PARK, CALIFORNIA ♦ LONDON ♦ AMSTERDAM ♦ DON MILLS, ONTARIO ♦ SYDNEY

Sponsoring Editor	Bob Rogers
Production Editor	Marcia Mirski
Copy Editor	James K. Madru
Text Designer	Vanessa Piñeiro
Illustrator	Illustration Concepts, Michael Ockler
Cover Designer and Illustrator	Hannus Design Associates, Richard Hannus
Art Coordinator	Kristin Belanger
Production Manager	Karen M. Guardino
Production Coordinator	Peter Petraitis

The text of this book was composed in Trump by York Graphic Services.

Illustrations rendered and copyrighted by Irving Geis: Figures 1.1, 1.2, 1.3, 1.6, 1.7, 1.8, 1.9, 1.10, 1.11, 1.13, 1.14, 1.15, 1.16, 1.17, 1.18, 1.19, 1.20, 1.21, 1.23, 1.26, 1.27, 1.28, 1.29, 1.32, 1.33, 3.4, 3.7, 3.16, 3.17(a), 3.43, 3.44, 3.45, 3.46, 3.47, 3.54(a), 3.58(lower half), 3.59, 4.7, 4.8, 4.9, 4.10, 4.15, 4.21, 10.9, 12 opener, 18.16, 18.35(b).

Library of Congress Cataloging in Publication Data

Zubay, Geoffrey L.
Biochemistry.

Includes bibliographies and index.

1. Biological chemistry. I. Title.

QP514.2.Z83 1983 574.19'2 82-18502
ISBN 0-201-09091-0

Copyright © 1983 by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada.

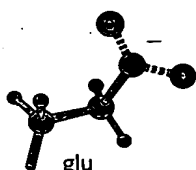
ISBN 0-201-09091-0
ABCDEFGHIJ-DO-89876543

EXTERNAL

ACIDIC

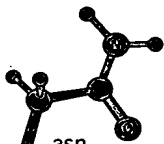


asp
aspartic acid (D)

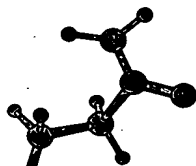


glu
glutamic acid (E)

NEUTRAL

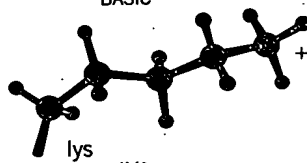


asn
asparagine (N)

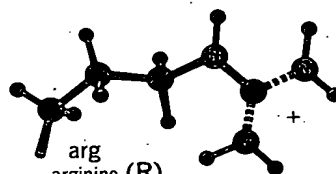


gln
glutamine (Q)

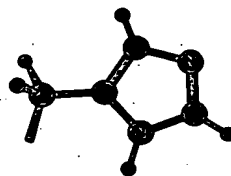
BASIC



lys
lysine (K)

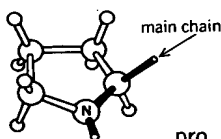


arg
arginine (R)

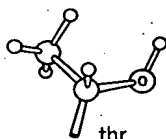


his
histidine (H)

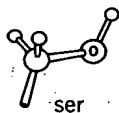
AMBIVALENT



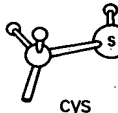
pro
proline (P)



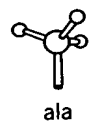
thr
threonine (T)



ser
serine (S)



cys
cysteine (C)

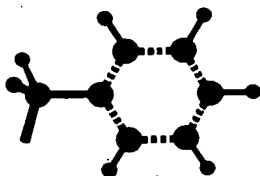


ala
alanine (A)



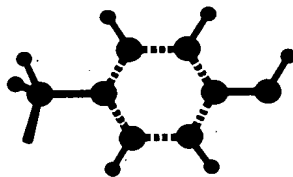
gly
glycine (G)

INTERNAL

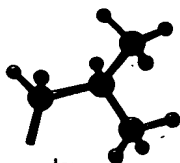
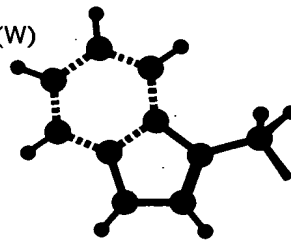


phe
phenylalanine (F)

tyr
tyrosine (Y)



trp
tryptophan (W)



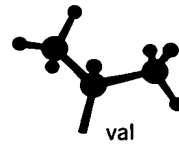
leu
leucine (L)



ile
isoleucine (I)



met
methionine (M)



val
valine (V)

include the small glycine (a single hydrogen atom) and alanine, serine and threonine (with attached hydroxyls), and cysteine (with its sulfhydryl). Proline has a hydrocarbon side chain, but its conformational properties put it at corners and therefore often outside.

Results of x-ray crystallography show these classifications by polarity and location to be valid in general for soluble globular proteins. The structures of myoglobin and hemoglobin, lysozyme, and cytochrome *c* all have buried hydrophobic side chains with hydrophilic side chains on the surface. Figure 1-11 shows the positions of all 104 side chains for horse heart cytochrome *c*. This is a protein with a heme group like myoglobin, but with an entirely different function. It is one of a chain of molecules that transports electrons in the mitochondria. Hydrophobic side chains (colored) pack inside the molecule, especially against the left side of the heme ring, and hydrophilic side chains (grey) are distributed over the surface of the molecule. This is a clear example of one way in which sequence dictates folding.

Other side chains have pronounced effects on three-dimensional conformation, particularly proline and the sulfur-containing cysteine. The side chain of proline contains a portion of the main chain and thus tends to change the direction of the main chain. Proline is often used to produce a bend in the protein chain, and many of the α helices in myoglobin and hemoglobin begin with a proline residue. The side chain —SH of cysteine can make a covalent —S—S— linkage with a similar residue from another protein chain (Figure 1-12). After the protein chain has reached its optimal low-energy conformation, the disulfide bonds can increase its stability. The enzyme ribonuclease contains four such disulfide bridges. If the —S—S— linkages are broken and the protein chain is made to unfold in the presence of a denaturing agent, such as urea, would it refold when the denaturing chemicals were removed? Christian Anfinsen and coworkers answered this question in the affirmative in the early 1960s with a classic set of experiments.

We have seen that sequence determines folding, but, in fact, it does more than that. It determines a unique folding pattern. The importance of the folding pattern can be appreciated through a consideration of the protein's function. Enzymes, for example, are molecular machines that operate with great precision on other molecules called substrates. Chymotrypsin is one of a class of pancreatic digestive enzymes that cuts other protein chains. The substrate is a polypeptide chain that is held on the surface of the enzyme so that a peptide bond can be cleaved. It is necessary that the substrate mesh with the enzyme in an exact lock-and-key fashion. In chymotrypsin there is a specificity pocket that fits an aromatic ring side chain of the substrate. Immediately adjoining the specificity pocket is an active site that assists in cutting a peptide bond near the bound aromatic ring.

◀ Figure 1-10

The 20 amino acid side chains classified by their probable position in the protein molecule. Three-letter and one-letter codes are given for each. The forms shown here are the most prevalent at pH 7. Note that histidine can play a dual role—neutral (as shown here) or positively charged.

Exhibit 46



UNITED STATES PATENT AND TRADEMARK OFFICE

500
UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/776,191	02/02/2001	Edwin L. Madison	24745-1607	3237
20985	7590	04/21/2006	EXAMINER	
FISH & RICHARDSON, PC P.O. BOX 1022 MINNEAPOLIS, MN 55440-1022			PAK, YONG D	
			ART UNIT	PAPER NUMBER
			1652	

DATE MAILED: 04/21/2006

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No. 09/776,191	Applicant(s) MADISON ET AL	
	Examiner Yong D. Pak	Art Unit 1652	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --
Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 30 January 2006.
 2a) ☐ This action is FINAL. 2b) ☒ This action is non-final.
 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) See Continuation Sheet is/are pending in the application.
 4a) Of the above claim(s) 1-3, 5, 10-13, 19-20, 34-36, 40-46, 48-55, 108-109, 113-116, 118-120 and 122-126
 is/are withdrawn from consideration.
 5) ☐ Claim(s) _____ is/are allowed.
 6) ☒ Claim(s) 1-3, 5, 11-13, 19, 20, 34-36, 40-42, 113 and 114 is/are rejected.
 7) ☐ Claim(s) _____ is/are objected to.
 8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☐ The specification is objected to by the Examiner.
 10) ☐ The drawing(s) filed on _____ is/are: a) ☐ accepted or b) ☐ objected to by the Examiner.
 Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
 Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
 a) ☐ All b) ☐ Some * c) ☐ None of:
 1. ☐ Certified copies of the priority documents have been received.
 2. ☐ Certified copies of the priority documents have been received in Application No. _____.
 3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).
 * See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- | | |
|--|---|
| 1) <input type="checkbox"/> Notice of References Cited (PTO-892) | 4) <input type="checkbox"/> Interview Summary (PTO-413)
Paper No(s)/Mail Date. _____ |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948) | 5) <input type="checkbox"/> Notice of Informal Patent Application (PTO-152) |
| 3) <input type="checkbox"/> Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08)
Paper No(s)/Mail Date _____ | 6) <input type="checkbox"/> Other: _____ |

Continuation Sheet (PTOL-326)

Application No. 09/776,191

Continuation of Disposition of Claims: Claims pending in the application are 1-3,5,10-13,19,20,34-36,40-46,48-55,108,109,113-116,118-120 and 122-126.

DETAILED ACTION

This application is a CIP of 09/657,986, now issued as U.S. Patent No. 6,797,504.

Continued Examination Under 37 CFR 1.114

A request for continued examination under 37 CFR 1.114, including the fee set forth in 37 CFR 1.17(e), was filed in this application after final rejection. Since this application is eligible for continued examination under 37 CFR 1.114, and the fee set forth in 37 CFR 1.17(e) has been timely paid, the finality of the previous Office action has been withdrawn pursuant to 37 CFR 1.114. Applicant's submission filed on January 30, 2006, amending claims 1, 5, 12-13, and 113-114 and canceling claims 6-7, 9-10, 14, 16, 18 and 137, has been entered.

Claims 1-3, 5, 10-13, 19-20, 34-36, 40-46, 48-55, 108-109 113-116, 118-120 and 122-126 are pending. Claims 1-3, 5, 10-13, 19-20, 34-36, 40-46, 48-55, 108-109 113-116, 118-120 and 122-126 are withdrawn. Claims 1-3, 5, 11-13, 19-20, 34-36, 40-42 and 113-114 are under consideration.

Priority

Applicant's claim for domestic priority under 35 U.S.C. 119(e) is acknowledged. However, the provisional applications upon which priority is claimed fails to provide adequate support under 35 U.S.C. 112 for claims 11-13 and 34 of this application.

Provisional applications 60/179,982, 60/183,542, 60/213,124, 60/220,970 and 60/234,840 fail to provide adequate support for polypeptides comprising the serine protease domain of MTSP1. Provisional applications 60/179,982 and 60/183,542 describe polypeptides related MTSP3 and provisional application 60/213,124, 60/220,970 and 60/234,840 describe polypeptides related to MTSP4.

Therefore, the effective filing date for purpose of prior art is the filing date of 09/657,986, which is 9/8/2000.

Response to Arguments

Applicant's amendment and arguments filed on January 30, 2006, have been fully considered and are deemed to be persuasive to overcome the rejections previously applied. Rejections and/or objections not reiterated from previous office actions are hereby withdrawn.

Claim Objections

Claims 11-13 and 34 are objected for being drawn to non-elected subject matter. In response to the previous Office Action, applicants have traversed the above rejection. Applicants argue that claims 11-13 and 34 are directed to elected subject matter. Even though claims are drawn to MTSP1, the elected subject matter, the claims are also drawn to non-elected subject matter, i.e. MTSP3 (SEQ ID NO:4), MTSP4 (SEQ DI NO:6), MTSP6 (SEQ DI NO:12), corin, enteropeptidase, human airway trypsin-like protease , TMPRSS2, TMPRSS4. Hence the objection is maintained.

Claim Rejections - 35 USC § 112

The following is a quotation of the second paragraph of 35 U.S.C. 112:

The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

Claims 1-3, 5, 11-12, 13 and claims 19-20, 34-36, 40-42 and 113-114 depending therefrom rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claims 1-3, 5, 11-12, 13 recite the phrase "substantially purified single-chain polypeptide". The metes and bounds of the phrase in the context of the above claims are not clear to the Examiner. It is not clear to the Examiner what is considered as "substantially purified" by the applicants. A perusal of the specification did not provide a clear definition for the above phrase. Without a clear definition, those skilled in the art would be unable to conclude if a polypeptide is a "substantially purified" polypeptide without knowing the metes and bounds of the phrase. Examiner requests clarification of the above phrase.

Claim 1 and claims 2-3, 5, 11-13, 19-20, 34-36, 40-42 and 113-114 depending therefrom are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claim 1 recites the phrase "the MTSP protease domain or catalytically active fragment thereof is the only portion of the single-chain polypeptide from the MTSP".

The metes and bounds of the phrase in the context of the claim is not clear. It is not clear to the Examiner as to how one skilled in the art would identify a given amino acid sequence as being "from MTSP" or not being "from MTSP". Examiner has interpreted the claims broadly to mean that a "single-chain polypeptide comprising a MTSP protease domain or catalytically active fragment thereof is the only portion of the single-chain polypeptide from the MTSP" is a "single-chain polypeptide comprising a fragment consisting of a protease domain or a catalytically active fragment thereof". Examiner requests clarification of the above phrase.

Claims 12-13 and claims 113-114 depending therefrom are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claims 12-13 recite the phrase "protease domain has a sequence of amino acid residues set forth as amino acids 615-855 of SEQ ID NO:2" or "protease domain whose sequence of amino acid residues is set forth as amino acid residues 615-855 of SEQ ID NO:2". The metes and bounds of the phrase in the context of the claims are not clear. It is not clear to the Examiner if the recited amino acid sequence has the amino acid sequence of SEQ ID NO:2 or is a representative member of a genus. Examiner suggests amending the phrase as "protease domain comprises amino acids 615-855 of SEQ ID NO:2" to clearly indicate that the protease domain has the amino acids 615-855 of SEQ ID NO:2.

Claim 19-20 are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claims 19-20 recite the phrase "free Cys". The metes and bounds of the phrase in the context of the above claims are not clear to the Examiner. It is not clear to the Examiner what is considered as "free Cys" by the applicants. A perusal of the specification did not provide a clear definition for the above phrase. Without a clear definition, those skilled in the art would be unable to conclude if Cys is "free". Examiner requests clarification of the above phrase.

Claim 19 is rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claim 19 recites the phrase "exhibits proteolytic activity". The metes and bounds of the phrase in the context of the above claim are not clear to the Examiner. It is not clear to the Examiner either from the specification or from the claims as to what applicants mean by the above phrase. Examiner requests clarification of the above phrase.

The following is a quotation of the first paragraph of 35 U.S.C. 112:

The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

Claims 1-3, 5, 9, 11, 19-20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

Claims 1-3, 5, 9, 19-20, 35-36, 40-42 and 113-114 are drawn to a polypeptide comprising a protease or catalytically active portion of type-II membrane-type serine protease (MTSP) from any source. Claims 11 and 34 limit the MTSP polypeptide to a MTSP1 polypeptide from any source. Therefore, these claims are drawn to a genus of polypeptides having any structure. The specification only teaches four species, amino acids 615-855 of SEQ ID NO:2, amino acids of 205-437 of SEQ ID NO:4, amino acids of SEQ ID NO:6 and amino acids 217-443 of SEQ ID NO:11. These species are not enough to describe the whole genus and there is no evidence on the record of the relationship between the structure of the above catalytically active protease domains of SEQ ID NOs: 2, 4, 6 and 11 and the structure of the serine protease domain of any or all MTSP polypeptides or MTSP1 polypeptides. Further, the specification does not describe the structure of a catalytically active portion of any or all MTSP polypeptide. Therefore, the specification fails to describe a representative species of the genus of polypeptides comprising of a serine protease domain or a catalytically active portion of a MTSP polypeptide.

Given this lack of description of the representative species encompassed by the genus of the claims, the specification fails to sufficiently describe the claimed invention

in such full, clear, concise, and exact terms that a skilled artisan would recognize that applicants were in possession of the inventions of claims 1-3, 5, 9, 11, 19-20, 34-36, 40-42 and 113-114.

Applicant is referred to the revised guidelines concerning compliance with the written description requirement of U.S.C. 112, first paragraph, published in the Official Gazette and also available at www.uspto.gov.

In response to the previous Office Action, applicants have traversed the above rejection.

Applicants argue that the claims meet the written description guideline since the specification teaches common elements of MTSP and protease domains of MTSPs, thereby providing structural and functional characteristics of the various species. Applicants also argue that the specification explicitly provides several catalytically active portions of MTSP, SEQ ID NO:2, 4, 6 and 11 (MTSP1, MTSP3, MTSP4 and MTSP 6), along with how to make other catalytically active fragments of MTSP, and therefore, the specification provides "relevant, identifying characteristics" of a representative number of species of the claimed genus. Examiner respectfully disagrees. The claims are drawn to polypeptides comprising any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1. The claims are drawn to polypeptides having any structure and therefore, the claims are drawn to a genus encompassing species having substantial variation and fails to describe a representative number of species. As discussed in the written description guidelines,

the written description requirement for a claimed genus may be satisfied through sufficient description of a representative number of species by actual reduction to practice, reduction to drawings, or by disclosure of relevant, identifying characteristics, i.e., structure or other physical and/or chemical properties, by functional characteristics coupled with a known or disclosed correlation between function and structure, or by a combination of such identifying characteristics, sufficient to show the applicant was in possession of the claimed genus. A representative number of species means that the species which are adequately described are representative of the entire genus. **Thus, when there is substantial variation within the genus, one must describe a sufficient variety of species to reflect the variation within the genus.** Satisfactory disclosure of a representative number depends on whether one of skill in the art would recognize that the applicant was in possession of the necessary common attributes or features of the elements possessed by the members of the genus in view of the species disclosed. For inventions in an unpredictable art, adequate written description of a genus which embraces widely variant species cannot be achieved by disclosing only one species within the genus. In the instant case the claimed genera of the claims are drawn to species which are widely variant in structure. The genus of the claims are structurally diverse as it encompasses any catalytically active protease domains of any or all MTSP or MTSP1, excepting having serine protease activity. As such, neither the description of solely structural features present in all members of the genus is sufficient to be representative of the attributes and features of the entire genus.

Hence the rejection is maintained.

Claims 1-3, 5, 9, 19-20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 112, first paragraph, because the specification, while being enabling for a polypeptide comprising amino acids 615-855 of SEQ ID NO:2, does not reasonably provide enablement for a polypeptide comprising any protease domain of any type II membrane type serine protease (MTSP) or MTSP1 or a catalytically active portion thereof. The specification does not enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the invention commensurate in scope with these claims.

Factors to be considered in determining whether undue experimentation is required are summarized in In re Wands 858 F.2d 731, 8 USPQ2nd 1400 (Fed. Cir. 1988). They include (1) the quantity of experimentation necessary, (2) the amount of direction or guidance presented, (3) the presence or absence of working examples, (4) the nature of the invention, (5) the state of the prior art, (6) the relative skill of those in the art, (7) the predictability or unpredictability of the art, and (8) the breadth of the claims.

Claims 1-3, 5, 9, 19-20, 35-36, 40-42 and 113-114 are drawn to a polypeptide comprising a protease or catalytically active portion of type-II membrane-type serine protease (MTSP) from any source. Claims 11 and 34 limit the MTSP polypeptide to a MTSP1 polypeptide from any source. Therefore, these claims are drawn to polypeptides having undefined structure.

The scope of the claims is not commensurate with the enablement provided by the disclosure with regard to the extremely large number of polypeptides comprising a

protease or catalytically active domain broadly encompassed by the claims. Since the amino acid sequence of a protein determines its structural and functional properties, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a knowledge of and guidance with regard to which amino acids in the protein's sequence, if any, are tolerant of modification and which are conserved (i.e. expectedly intolerant to modification), and detailed knowledge of the ways in which the proteins' structure relates to its function. However, in this case the disclosure is limited to the polypeptide comprising amino acids 615-855 of SEQ ID NO:2, or the amino acids of SEQ ID NO:50.

It would require undue experimentation of the skilled artisan to make and use the claimed polypeptides. The specification is limited to teaching the use of polypeptide comprising amino acids 615-855 of SEQ ID NO:2 or the amino acids of SEQ ID NO:50 but provides no guidance with regard to the making of variants and mutants or with regard to other uses. In view of the great breadth of the claim, amount of experimentation required to make the claimed polypeptides, the lack of guidance, working examples, and unpredictability of the art in predicting function from a polypeptide primary structure, the claimed invention would require undue experimentation. As such, the specification fails to teach one of ordinary skill how to use the full scope of the polypeptides encompassed by the claims.

While enzyme isolation techniques, recombinant and mutagenesis techniques are known, and it is routine in the art to screen for multiple substitutions or multiple modifications as encompassed by the instant claims, the specific amino acid positions

within a protein's sequence where amino acid modifications can be made with a reasonable expectation of success in obtaining the desired activity/utility are limited in any protein and the result of such modifications is unpredictable. In addition, one skilled in the art would expect any tolerance to modification for a given protein to diminish with each further and additional modification, e.g. multiple substitutions.

The specification does not support the broad scope of the claims which encompass all modifications and variants of a protease or catalytically active domain or modifications of amino acids 615-855 of SEQ ID NO:2 because the specification does not establish: (A) regions of the protein structure which may be modified without affecting MTSP/serine protease activity; (B) the general tolerance of MTSP to modification and extent of such tolerance; (C) a rational and predictable scheme for modifying any amino acid residue with an expectation of obtaining the desired biological function; and (D) the specification provides insufficient guidance as to which of the essentially infinite possible choices is likely to be successful.

Thus, applicants have not provided sufficient guidance to enable one of ordinary skill in the art to make and use the claimed invention in a manner reasonably correlated with the scope of the claims broadly including protease or catalytically active domains of MTSP with an enormous number of amino acid modifications of the MTSP polypeptides and of amino acids 615-855 of SEQ ID NO:2. The scope of the claims must bear a reasonable correlation with the scope of enablement (*In re Fisher*, 166 USPQ 19 24 (CCPA 1970)). Without sufficient guidance, determination of the serine protease domain or the catalytically active domain of MTSP having the desired biological

characteristics is unpredictable and the experimentation left to those skilled in the art is unnecessarily, and improperly, extensive and undue. See *In re Wands* 858 F.2d 731, 8 USPQ2nd 1400 (Fed. Cir, 1988).

In response to the previous Office Action, applicants have traversed the above rejection.

Applicants argue that the level of skill in this art is high and the specification teaches structural and functional features sufficient to enable one of skill in the art to make sue the single chain polypeptides comprising catalytically active portion of an MTSP protease domain, by providing structure of MTSP polypeptides and their protease domains, as well as their conserved structures. Examiner respectfully disagrees. The scope of the claims, which are drawn to polypeptides comprising any protease domains or any or all catalytically active fragments of said protease domains of any or all MTSP or any or all MTSP1, including any or all recombinants, variants and mutants of said MTSP or MTSP1, is not commensurate with the enablement provided by the disclosure with regard to the extremely large number of polypeptides comprising a protease or catalytically active domain broadly encompassed by the claims. Even though the structure of some MTSP are known, the claims are drawn to any or all catalytically active fragments of any or all protease domains of any or all MTSP or MTSP1. As discussed above, predictability of which changes can be tolerated in a protein's amino acid sequence and obtain the desired activity requires a specific knowledge of and guidance with regard to which specific amino acids in the protein's sequence, can be modified such that the modified polypeptide continues to have said

claimed activity. It is this specific guidance that applicants do not provide. While the art may teach in general the structure of MTSP conserved amino acid sequences, protease domains, X-ray crystal structure and etc, such teachings will not reduce the burden of undue experimentation on those of ordinary skill in the art.

Applicants argue that the specification discloses working examples, thus a person skilled in the art has sufficient guide in making the claimed polypeptides. Examiner respectfully disagrees. Even though the structure of some MTSP are taught, the claims are not only drawn to polypeptides comprising catalytically active fragments of only MTSP1, MTSP3, MTSP4 and MTSP6, but to any or all mutants, variants and recombinants of any MTSP. Without specific guidance, those skilled in the art will be subjected to undue experimentation of making and testing each of the enormously large number of mutants that results from such experimentation. While the art may teach in general the structure of MTSP, conserved amino acid sequences, and etc, such teachings will not reduce the burden of undue experimentation on those of ordinary skill in the art.

Hence the rejection is maintained.

Applicants argue that it would be unfair, unduly limiting and contrary to the public policy upon which the patent laws are based to require applicant to limit the instant claims to only one exemplified protease domain. This argument is moot since patentability is based on statutes under 35 USC 101, 112, 102 and/or 103.

Claim Rejections - 35 USC § 102

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

- (a) the invention was known or used by others in this country, or patented or described in a printed publication in this or a foreign country, before the invention thereof by the applicant for a patent
- (b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.
- (e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

Claims 1-3, 5, 11-13, 19-20, 34-36, 40-42 and 113-114 are rejected under 35 U.S.C. 102(b) as being anticipated by Takeuchi et al. (see rejection of the phrase "MTSP protease domain or catalytically active fragment there is the only portion of the single-chain polypeptide from the MTSP" under 35 USC 112, 2nd paragraph above)

Claims 1-3, 5, 11-13, 19-20 and 34 are drawn to a polypeptide comprising fragment consisting of a serine protease domain of MTSP having the characteristics recited in the claims. Claims 35-36 are drawn to a conjugate comprising a polypeptide comprising a serine protease domain of MTSP and a targeting agent. Claims 40 –42 and 113-114 are drawn to a solid support comprising a polypeptide comprising a serine protease domain of MTSP.

Takeuchi et al. (Reference IJ : PTO-1449) teaches a polypeptide comprising a fragment consisting of a serine protease domain that is 100% identical to amino acids 615-855 of SEQ ID NO:2 of the instant invention (page 11060, 2nd full paragraph). Takeuchi et al. discloses a purified activated protease domain, comprising amino acids

615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VVGGT sequence (Figure 3 and right column on page 11057). The MTSP of Takeuchi et al. is not expressed on normal endothelial cells (page 11054, last paragraph and page 11055, 2nd full paragraph), is of human origin (Figure 1), consists essentially of the protease domain having catalytic activity (page 11060, 2nd full paragraph), and is expressed in tumor cells (page 11055, top paragraph).

Takeuchi et al. teaches a catalytically active polypeptide comprising the serine protease domain linked to a His-tag (page 11055, 3rd full paragraph, page 11057, 4th full paragraph). Takeuchi et al. also teaches a solid support comprising said polypeptide (page 11057, 4th full paragraph and Figure 5). Therefore, the teaching of Takeuchi et al. anticipates claims 1-3, 5, 11-13, 19-20, 34-36, 40-42 and 113-114 are.

Examiner notes that the contents of the reference were made public at the National Academy of Sciences colloquium held February 20-21, 1999 (see top of reference).

In response to the previous Office Action, applicants have traversed the above rejections.

Applicants argue that Takeuchi et al. does not anticipate the instant claims because it fails to disclose any polypeptides that incorporate all the features of claim 1, a single chain polypeptide having an MTSP portion, wherein the MTSP portion is a protease domain or a smaller fragment and wherein the MTSP portion has serine protease activity.

Applicants argue that the MT-SP1 of Takeuchi et al. is a full-length protein that includes additional MTSP regions other than a protease domain, and therefore, said MTSP1 of Takeuchi et al. is not a polypeptide where the only MTSP portion of the polypeptide is a protease domain or a smaller catalytically active portion of the protease domain. Examiner respectfully disagrees. First, the claim recites "a polypeptide comprising a MTSP portion" and the claim does not recite the limitation that the polypeptide only consist of MTSP portion. Therefore, a full-length MT-SP1 of Takeuchi et al. anticipates the instant claims. Second, in addition to the full-length MT-SP1, Takeuchi et al. also discloses a purified activated protease domain, comprising amino acids 615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VGGT sequence (Figure 3 and right column on page 11057). Even applicants state that Takeuchi et al. discloses "that its protease domain has an amino acid sequence containing amino acids 615-855 (Remarks page 36) and that "Takeuchi et al. discloses that its polypeptide includes the pro-domain and that the pro-domain is cleaved during auto-activation, resulting in a protease domain" (page 37). Therefore, said purified, activated protease domain anticipates the instant claims.

Applicants also argue that the reference of Takeuchi et al. does not anticipate the instant claims because the "purified protease domain" of Takeuchi et al. includes the His-tag sequence and that the polypeptide construct disclosed by Takeuchi et al. includes a sequence of 19 amino acids of a portion of the pro-domain and that his pro-domain is disulfide bonded to the protease domain. Examiner respectfully disagrees.

Takeuchi et al. also discloses a purified activated protease domain, comprising amino acids 615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VGGT sequence (Figure 3 and right column on page 11057 and Figure 6). Further, applicants state that "Takeuchi et al. discloses that its polypeptide includes the pro-domain and that the pro-domain is cleaved during auto-activation, resulting in a protease domain" (page 37).

Applicants also argue that the activated protein derived from the expressed His-tag amino acids 596-855 of MT-SP1 of Takeuchi et al. is not a single chain polypeptide because the protease domain is disulfide bonded to a pro-domain resulting in a two chain form. Examiner respectfully disagrees. Takeuchi et al. discloses that the pro-domain is disulfide bonded to a protease domain of the full length protein. Contrary to applicants argument, Takeuchi et al. does not teach that the pro-domain is disulfide bonded to an activated protease domain. Further, a single chain polypeptide is one sequence of amino acids beginning with a carboxyl end and terminating with an amino end, wherein the amino acids are connected via peptide bonds. Therefore, even the full length MT-SP1 of Takeuchi et al. having disulfide bonds can be construed as a single chain polypeptide.

In conclusion, Takeuchi et al. discloses a purified activated protease domain, comprising amino acids 615-855 of SEQ ID NO:2, confirmed by an N-terminal sequence of the purified, activated protease domain yielding the expected VGGT sequence (Figure 3 and right column on page 11057 and Figure 6). Further, applicants state that

"Takeuchi et al. discloses that its polypeptide includes the pro-domain and that the pro-domain is cleaved during auto-activation, resulting in a protease domain" (page 37).

Hence the rejections are maintained.

Claim Rejections - 35 USC § 102/103

The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

A person shall be entitled to a patent unless –

(e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

The following is a quotation of 35 U.S.C. 103(a), which forms the basis for all obviousness rejections, set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

Claims 1-3, 5, 10-13 and 34 rejected under 35 U.S.C. 102(e) as anticipated by or, in the alternative, under 35 U.S.C. 103(a) as obvious over O'Brien et al.

Claims 1-3, 5, 10-13 and 34 are drawn to a polypeptide comprising a serine protease domain of MTSP.

O'Brien et al. (U.S. Patent No. 5,972,616 – reference P- PTO 1449) teaches a polypeptide having 100% identity to the full length MTSP1 of SEQ ID NO:2 of the instant invention (SEQ ID NO:2, columns 19-24). The properties recited in claims 2-3 are

inherent properties of MTSP1 taught by O'Brien et al. since the polypeptide of O'Brien et al. and the instant invention have identical structure and therefore identical properties.

O'Brien et al. teaches a serine protease domain having proteolytic activity that is 100% identical to amino acids 615-855 of SEQ ID NO:2 (Figure 2, Figure 10 and SEQ ID NO:14). Although the protease domain of O'Brien et al. identified by SEQ ID NO:14 has not been purified, the protease domain in the reference and the polypeptide claimed by the applicants are one and the same. Therefore, the protease domain anticipates the instant invention.

Since the Office does not have facilities for examining and comparing applicant's polypeptide with the polypeptide of the prior art, the burden is on the applicant to show a novel or unobvious difference between the claimed product and the product of the prior art (i.e., that the polypeptide of the prior art does not possess the same material structure and functional characteristics of the claimed polypeptide). See *In re Best*, 562 F.2d 1252, 195 USPQ 430 (CCPA 1977) and *In re Fitzgerald et al.*, 205 USPQ 594.

Alternatively, O'Brien et al. teaches a method of expressing polypeptides via a vector in host cells. O'Brien et al. also teaches that the protease domain could be released and used as a diagnostic which has the potential for a target for therapeutic intervention (Column 15, lines 35-38). Therefore, it would have been obvious to one having ordinary skill in the art at the time the invention was made to express the protease domain of SQ ID NO:14 and purify the polypeptide. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for

Art Unit: 1652

therapeutic intervention. One of ordinary skill in the art would have had a reasonable expectation of success since expression of a heterologous polypeptide is routine in the art and O'Brien et al. teaches how to express heterologous polypeptides.

In response to the previous Office Action, applicants have traversed the above rejections.

Applicants argue that O'Brien et al. does not anticipate any of the instant claims because the claims are not directed to a full-length MTSP polypeptide. Examiner respectfully disagrees. The claim recites "a polypeptide comprising a MTSP portion" and the claim does not recite the limitation that the polypeptide only consist of MTSP portion. Therefore, the full-length MT-SP1 of O'Brien et al. anticipates the instant claims.

Applicants also argue that one of skill in the art would recognize the disclosure of the polypeptide of O'Brien as not disclosing a single chain polypeptide. Examiner respectfully disagrees. A single chain polypeptide is one sequence of amino acids beginning with a carboxyl end and terminating with an amino end, wherein the amino acids are connected via peptide bonds. Therefore, the full length MT-SP1 of O'Brien et al. can be construed as a single chain polypeptide.

Applicants argue that one of skill in the art would understand MTSP serine proteases to be active only as two chain polypeptides by citing Lu et al. (1999) *J. Biol. Chem.* 272:31293-300 and would not view O'Brien et al. as disclosing a single chain polypeptide. Examiner respectfully disagrees. The bibliographi information Lu et al. (1999) *J. Biol. Chem.* 272:31293-300 could not be located through *J. Biol. Chem.*

Applicants are urged to supply the reference or the correct bibliographic information. Nevertheless, applicants state that "as expressed, the MTSP polypeptide is an inactive single-chain zymogen" (Remarks page 42). Therefore, according to applicants, the full length MT-SP1 of O'Brien et al. is a single chain polypeptide and therefore, anticipates the claimed invention.

Hence the rejection is maintained.

Applicants also argue that O'Brien et al. provides no teaching or suggestion of smaller fragments having serine protease activity because it does not teach how to make a single chain polypeptide that has serine protease activity. Examiner respectfully disagrees. O'Brien et al. teaches a method of expressing polypeptides via a vector in host cells. It is well within the skill available in the art to purify the protease domain since O'Brien et al. identifies the protease domain. Therefore, it would have been obvious to one having ordinary skill in the art at the time the invention was made to express the protease domain of SQ ID NO:14 and purify the polypeptide. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. One of ordinary skill in the art would have had a reasonable expectation of success since expression of a heterologous polypeptide is routine in the art and O'Brien et al. teaches how to express heterologous polypeptides.

Applicants again argue that at the time of filing the instant application, one of skill in the art would not have had a reasonable expectation of success to express the

protease domain because art evidences that a single-chained polypeptide would not have been expected to have protease activity. Examiner respectfully disagrees. The claims are drawn to a polypeptide comprising a fragment consisting of a protease domain of SEQ ID NO:2. Therefore, said polypeptide being a single-chained polypeptide is an inherence property of said polypeptide since two polypeptides having identical structure will have identical function and physical and chemical properties.

Hence the rejections are maintained.

Claims 35-36, 40-42 and 113-114 are rejected under 35 U.S.C. 103(a) as being unpatentable over O'Brien et al.

Claims 35-36 are drawn to a conjugate comprising a polypeptide comprising a serine protease domain of MTSP and a targeting agent. Claims 40-42 and 113-114 are drawn to a solid support comprising a polypeptide comprising a serine protease domain of MTSP.

O'Brien et al. (U.S. Patent No. 5,972,616 – reference P- PTO 1449) teaches a polypeptide having 100% identity to the full length MTSP1 of SEQ ID NO:2 of the instant invention, as discussed above. O'Brien et al. also teaches that the protease domain could be released the used as a diagnostic which has the potential for a target for therapeutic intervention (Column 15, lines 35-38).

O'Brien et al. also teaches method of making fragments of SEQ ID NO:2 (Column 9, lines 22-55). O'Brien et al. teaches said fragments linked to another polypeptide (Column 9, lines 54-55) and conjugated to bridging molecules (Column 6,

lines 27-39) for detecting the polypeptide. Assays using polypeptides linked to the molecules taught by O'Brien et al. utilize solid supports.

Therefore, it would have been obvious to one having ordinary skill in the art at the time the claimed invention was made to make a polypeptide comprising of the serine protease domain of SEQ ID NO:2 taught by O'Brien et al. and to make conjugates and solid support comprising of a polypeptide comprised of the serine protease domain of SEQ ID NO:2. The motivation of making such a polypeptides is to use it as a diagnostic which has the potential for a target for therapeutic intervention. The motivation of making conjugates and solid supports comprising of said polypeptide is to use the conjugate and solid support in a variety of diagnostic assays. One of ordinary skill in the art would have had a reasonable expectation of success making fragments of a polypeptide is routine in the art and O'Brien et al. teaches how to make fragments of SEQ ID NO:2. One of ordinary skill in the art would have had a reasonable expectation of success in diagnostic assays using conjugates and solid supports comprising a polypeptide is very well known, as taught by O'Brien et al.

Therefore, the above references render claims 35-36 and 40-42 *prima facie* obvious to one of ordinary skill in the art.

In response to the previous Office Action, applicants have traversed the above rejections. Applicants argue that the teachings of O'Brien et al. does not result in the instantly claimed compositions because O'Brien et al. does not teach or suggest a single chain polypeptide that includes a MTSP protease domain where the polypeptide

does not include any additional MTSP portions and the polypeptide has serine protease activity. O'Brien et al. does teach or suggest a single chain polypeptide comprising a MTSP portion, wherein the MTSP portion is a protease domain and wherein the MTSP portion has serine protease activity and wherein the MTSP portion is the only portion of the polypeptide because O'Brien et al. identifies the serine protease domain and one having ordinary skill in the art at the time the invention was filed would have been motivated to purify the serine protease domain of O'Brien et al. as discussed above.

Hence the rejection is maintained.

Claims 19-20 are rejected under 35 U.S.C. 103(a) as being unpatentable over O'Brien et al. and Estell et al. in view of Takeuchi et al.

Claims 19-20 are drawn to a polypeptide comprising the serine protease domain of a MTSP wherein free Cys residues are substituted with Ser residues.

O'Brien et al. teaches a serine protease domain of a MTSP polypeptide, as discussed above.

The reference of O'Brien et al. does not teach a serine protease domain of a MTPSP polypeptides wherein free Cys residues have been replaced with Ser residues.

It is well known in the art that proteins form disulfide bonds via the SH groups of Cys residues. Upon making a polypeptide comprising a serine protease domain, a Cys residue which normally forms disulfide bonds in the full length polypeptide may be left free. For example, Takeuchi et al. (Reference IJ : PTO-1449) teaches that Cysteine at

position 731 of SEQ ID NO:2 normally forms a disulfide bond with a Cys residue in the pro-protease domain (see page 11060, top left paragraph and Figures 1 and 2).

Cys residues are sensitive to oxidation due to their SH side group. Estell et al. (U.S. Patent No. 5,346,823) teaches that Cys residues replaced with Ser residues to decrease a polypeptide's susceptibility to oxidation (Abstract and Column 10, lines 34-38). Ser residues have similar side chains as Cys residues and substitution of a Cys residue with a Ser residue is a conservative substitution.

Therefore, it would have been obvious to one having ordinary skill in the art at the time the claimed invention was made to replace free Cys residues in the protease domain taught by O'Brien et al. with a Ser residue. One of ordinary skill in the art would be motivated to make such a change in order to enhance stability of the polypeptide. One of ordinary skill in the art would have had a reasonable expectation of success since Estell et al. teaches successful decrease of a protein's susceptibility to oxidation by substituting residues sensitive to oxidation with conservative substitutions.

Therefore, the above references render claims 1 and 16, 18-20, 34 and 137 *prima facie* obvious to one of ordinary skill in the art.

In response to the previous Office Action, applicants have traversed the above rejections. Applicants argue that the combination of the teachings of O'Brien et al. with the teachings of Estell et al., and Takeuchi et al. does not result in the instantly claimed methods because O'Brien et al. does not teach or suggest a single chain polypeptide that includes a MTSP protease domain where the polypeptide does not include any

Art Unit: 1652

additional MTSP portions and the polypeptide has serine protease activity and that neither Takeuchi et al. nor Estell et al. remedy the defects of O'Brien et al. First, the claims are product claims and not method claims. Second, O'Brien et al. does teach or suggest a single chain polypeptide comprising a MTSP portion, wherein the MTSP portion is a protease domain and wherein the MTSP portion has serine protease activity and wherein the MTSP portion is the only portion of the polypeptide because O'Brien et al. identifies the serine protease domain and one having ordinary skill in the art at the time the invention was filed would have been motivated to purify the serine protease domain of O'Brien et al. as discussed above.

Applicants argue that Takeuchi et al. teaches that every cysteine residue of the protein is disulfide bonded and therefore Takeuchi et al. does not teach or suggest an MTSP protease domain having a free Cys residue. Examiner respectfully disagrees. Figure 4 applicants are referring to illustrate disulfide bonds of cysteine residues of the full length MTSP, for example, the Cys at position 830 is disulfide bonded to Cys at position 191.

Hence the rejections are maintained.

None of the claims are in condition for allowance.

Application/Control Number: 09/776,191

Page 28

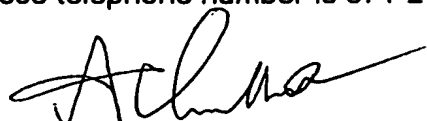
Art Unit: 1652

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Yong Pak whose telephone number is 571-272-0935. The examiner can normally be reached 6:30 A.M. to 5:00 P.M. Monday through Thursday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Ponnathapu Achutamurthy can be reached on 571-272-0928. The fax phone numbers for the organization where this application or proceeding is assigned are 571-273-8300 for regular communications and 703-872-9307 for After Final communications.

Any inquiry of a general nature or relating to the status of this application or proceeding should be directed to the receptionist whose telephone number is 571-272-1600.

Yong D. Pak
Patent Examiner 1652


PONNATHAPU ACHUTAMURTHY
SUPERVISORY PATENT EXAMINER
TECHNOLOGY CENTER 1600

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☒ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:**

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.